ICASE

THEORY OF SPECTRAL METHODS FOR MIXED INITIAL-BOUNDARY VALUE PROBLEMS PART I

DAVID GOTTLIEB

STEVEN A. ORSZAG

Report Number 76-32
November 26, 1976

INSTITUTE FOR COMPUTER APPLICATIONS IN SCIENCE AND ENGINEERING
NASA Langley Research Center, Hampton, Virginia

Operated by the

UNIVERSITIES SPACE USRA RESEARCH ASSOCIATION

THEORY OF SPECTRAL METHODS

FOR MIXED INITIAL-BOUNDARY VALUE PROBLEMS

PART I

David Gottlieb

Steven A. Orszag*

* Department of Mathematics, Massachusetts Institute of Technology

This paper was prepared as a result of work performed under NASA Contract Number NAS1-14101 while the first author was in residence at ICASE, NASA Langley Research Center, Hampton, VA 23665. The work performed by the second author was supported by the Office of Naval Research and the National Science Foundation.

1. Introduction

In this monograph we give a mathematical analysis of spectral methods for mixed initial-boundary value problems. This theory is also useful for analysis of a variety of finite element and finite difference methods (see Section 5). However, before proceeding to the formal presentation of the theory let us give some simple examples of the kinds of behavior we wish to explain.

Spectral methods involve representing the solution to a problem as a truncated series of known functions of the independent variables. We shall make this idea precise in Sec. 2, but we can illustrate it here by the standard separation of variables solution to the mixed initial-boundary value problem for the heat equation.

Example 1.1: Fourier sine series solution of the heat equation.

Consider the mixed initial-boundary value problem

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2} \qquad (0 < x < \pi,\ t \geq 0) \qquad (1.1a)$$

$$u(0,t) = u(\pi,t) = 0 \qquad (t > 0) \qquad (1.1b)$$

$$u(x,0) = f(x) \qquad (0 \leq x \leq \pi)\ . \qquad (1.1c)$$

The solution is

$$u(x,t) = \sum_{n=1}^{\infty} a_n(t) \sin n x\ , \qquad (1.2)$$

$$a_n(t) = f_n e^{-n^2 t} \quad (n=1,2,\ldots,) \tag{1.3}$$

where

$$f_n = \frac{2}{\pi}\int_0^{\pi} f(x) \sin nx \, dx \quad (n=1,2,\ldots,) \tag{1.4}$$

are the coefficients of the Fourier sine series expansion of $f(x)$. Recall that any function in $L_2(0,\pi)$ has a Fourier sine series that converges to it in $L_2(0,\pi)$; the Fourier sine series of any piecewise continuous function $f(x)$ which has bounded variation on $(0,\pi)$ converges to $\frac{1}{2}[f(x+)+f(x-)]$ throughout $(0,\pi)$ (see Section 3).

A spectral approximation is gotten by simply truncating (1.2) to

$$u_N(x,t) = \sum_{n=1}^{N} a_n(t) \sin nx \tag{1.5}$$

and replacing (1.3) by the evolution equation

$$\frac{da_n}{dt} = - n^2 a_n \qquad (n=1,\ldots,N) \; . \tag{1.6}$$

with the initial conditions $a_n(0) = f_n$ $(n=1,\ldots,N)$.

The spectral approximation (1.5-6) to (1.1) is an exceedingly good approximation for any $t > 0$ as $N \to \infty$. In fact, the error $u(x,t) - u_N(x,t)$ goes to zero more rapidly than $e^{-N^2 t}$ as $N \to \infty$ for any $t > 0$. In contrast, a finite difference approximation to the heat equation using N grid points

in x but leaving t as a continuous variable (a 'semi-discrete' approximation) leads to errors that decay only algebraically with N as $N \to \infty$. [Of course, if we solve (1.6) by finite differences in t the error of the spectral method would go to zero algebraically with the time step Δt. However, we shall neglect all time differencing errors for now and study only the convergence of semi-discrete approximations. Time-differencing methods are discussed in Section 10.]

Example 1.2: Fourier sine series solution of an inhomogeneous heat equation.

Not all spectral methods work as well as the trivial one just outlined in Example 1.1. Consider for example the solution to the problem

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + 1 \qquad (0 < x < \pi, \quad t \geq 0)$$

with the same initial and boundary conditions as before. The Fourier sine coefficients of the exact solution are now

$$a_n(t) = f_n\, e^{-n^2 t} + \frac{4}{\pi n^3}\,(1-e^{-n^2 t})\, e_n \qquad (1.7)$$

where $e_n = 0$ if n is even and $e_n = 1$ if n is odd. Spectral approximations are now given by (1.5) with (1.6) replaced by

$$\frac{da_n}{dt} = -\, n^2 a_n + \frac{4}{\pi n}\, e_n \qquad (n=1,\ldots,N)\ ,$$

the solution of which is (1.7) for $n = 1,\ldots,N$. Now the truncation error $u(x,t) - u_N(x,t)$ no longer decays exponentially as $N \to \infty$; the error is of order N^{-3} as $N \to \infty$ for fixed x, $0 < x < \pi$, and $t > 0$. In other words, the results to be anticipated from this spectral method behave asymptotically as $N \to \infty$ in the same way as those obtained by a third-order finite-difference scheme [in which the error goes to zero like $\Delta x^3 = (\pi/N)^3$].

The last example may be disturbing but even more serious difficulties confront the unwary user of spectral methods, as the next example should make amply clear.

Example 1.3: Fourier sine series solution of the one-dimensional wave equation.

Consider the mixed initial-boundary value problem for the one-dimensional wave equation

$$\frac{\partial u(x,t)}{\partial t} + \frac{\partial u(x,t)}{\partial x} = x + t \qquad (0 < x < \pi, \quad t \geq 0) \tag{1.8a}$$

$$u(0,t) = 0 \qquad (t \geq 0) \tag{1.8b}$$

$$u(x,0) = 0 \qquad (0 \leq x \leq \pi) \tag{1.8c}$$

The exact solution to this well posed problem is $u(x,t) = xt$. This solution can also be found by Fourier sine series expansion of $u(x,t)$. To do this, we substitute (1.2) into (1.8) and re-expand all terms in sine series. The Fourier expansion of $\partial u/\partial x$ is

$$\frac{\partial u}{\partial x} = \sum_{n=1}^{\infty} b_n(t) \sin nx \tag{1.09}$$

where integration by parts gives

$$b_n(t) = \frac{2}{\pi} \int_0^{\pi} \frac{\partial u}{\partial x} \sin nx \, dx, = - \frac{2n}{\pi} \int_0^{\pi} u \cos nx \, dx$$

$$= - \frac{2n}{\pi} \sum_{n=1}^{\infty} a_m(t) \int_0^{\pi} \sin mx \cos nx \, dx, \tag{1.10}$$

$$= \frac{4}{\pi} \sum_{\substack{m=1 \\ m+n \text{ odd}}}^{\infty} \frac{nm}{n^2-m^2} a_m(t) \, .$$

Also the Fourier sine coefficients of x are $2/n(-1)^{n+1}$ and the Fourier sine coefficients of t are $(4t/\pi n)e_n$, where $e_n = 0$ if n is even and $e_n = 1$ if n is odd. Equating coefficients of $\sin nx$ in (1.8a) we obtain

$$\frac{da_n}{dt} = - \frac{4}{\pi} \sum_{\substack{m=1 \\ m+n \text{ odd}}}^{\infty} \frac{nm}{n^2-m^2} a_m - \frac{2}{n} (-1)^n + \frac{4}{\pi n} t \, e_n \quad (n=1,2,\ldots). \tag{1.11}$$

The Fourier sine coefficients of the exact solution $u(x,t) = xt$ are

$$a_n(t) = - \frac{2}{n} (-1)^n t \qquad (n = 1,2,\ldots)$$

It is easy to verify by direct substitution that these coefficients satisfy (1.11) exactly; in particular, the sum in (1.11) converges for all t.

Now suppose we employ a spectral method based on Fourier sine series to solve this problem. We seek a solution to (1.8) in the form of the truncated sine series (1.4). If the exact coefficients $a_n(t)$ are used in (1.4) then $u(x,t) - u_N(x,t) \to 0$ as $N \to \infty$; for each fixed x , $0 < x < \pi$, and $t > 0$ the error is of order $1/N$ as $N \to \infty$ (see Section 3).

However, it is not reasonable to assume that the expansion coefficients $a_n(t)$ are known exactly in this case because of the complicated couplings between various n in the system (1.11). It is more reasonable to determine them by numerical solution of an approximation to (1.11). Galerkin approximation (see Sec. 2) gives the truncated system of equations

$$\frac{da_n}{dt} = -\frac{4}{\pi} \sum_{\substack{m=1 \\ m+n \text{ odd}}}^{N} \frac{nm}{n^2-m^2} a_m - \frac{2}{n}(-1)^n + \frac{4}{\pi n} te_n \quad (n=1,\ldots,N) \qquad (1.12)$$

The truncation of the infinite system (1.11) to the finite system (1.12) is a standard way to approximate infinite coupled systems. Unfortunately, it does not always work. In Figs. 1.1 - 1.2 we show plots of the approximations $u_N(x,t)$ at $t = 5$ given by (1.4) for $N = 50,75$. These plots are obtained by numerical solution of (1.12) with $a_n(0) = 0$; the time steps used in the numerical solution of (1.12) were so small that time differencing errors are negligible. It is apparent that the approximate solu-

tions with N finite do not converge to the exact solution as N increases. The divergence of this spectral method will be explained in Section 6.

Not all spectral methods give such poor results. A properly formulated and implemented spectral method gives results of striking accuracy with efficient use of computer resources.

The choice of an appropriate spectral method is governed by two main considerations:

(i) Accuracy. In order to be very useful a spectral method should be designed to give results of far greater accuracy than can be obtained by more conventional difference methods using similar spatial resolution or degrees of freedom. The choice of appropriate spectral representation depends on the kind of boundary conditions involved in the problem.

(ii) Efficiency. In order to be useful the spectral method should not be very much less efficient than difference methods with comparable numbers of degrees of freedom. For similar work, spectral methods should produce much more accurate results than conventional methods.

In Section 14, we present a catalog of different spectral methods and indicate the kinds of problems to which they can be most usefully applied.

Many examples of efficient and accurate spectral methods will be given later.

2. Spectral Methods

The problems to be studied here are mixed initial-boundary value problems of the form

$$\frac{\partial u(x,t)}{\partial t} = L(x,t)u(x,t) + f(x,t) \quad (x \in D\ ,\ t \geq 0) \tag{2.1}$$

$$B(x)u(x,t) = 0 \qquad (x \in \partial D,\ \ t > 0) \tag{2.2}$$

$$u(x,0) = g(x) \qquad (x \in D) \tag{2.3}$$

where D is a spatial domain with boundary ∂D, $L(x,t)$ is a linear (spatial) differential operator and $B(x)$ is a linear (time independent) boundary operator. Here we write (2.1-3) for a single dependent variable u and a single space coordinate x with the understanding that much of the following analysis generalizes to systems of equations in higher space dimensions. Also, attention is restricted to problems with homogeneous boundary conditions because the solution to any problem involving inhomogeneous boundary conditions is the sum of an arbitrary function having the imposed boundary values and a solution to a problem of the form (2.1-3). The extension to nonlinear problems will be indicated at the end of this section.

Before discussing spectral methods for solution of (2.1-3) let us set up the mathematical framework for our later analysis. It is assumed that, for each t, $u(x,t)$ is an element of a Hilbert space H with inner product $(\ ,\ )$ and norm $\|\cdot\|$. For each $t > 0$, the solution[1] $u(t)$ belongs to the subspace $\mathcal{B}$ of H consisting of all functions $u \in H$

[1] We will often denote $u(x,t)$ by $u(t)$ when discussing u as a function of t.

satisfying $Bu = 0$ on ∂D. We do not require that $u(x,0)=g(x) \in \mathcal{B}$ but only $u(x,0) \in \mathcal{H}$. The operator L is usually an unbounded differential operator whose domain is dense in $\mathcal{H}$ but does not include all functions $u \in \mathcal{H}$. For example, if $L = \partial/\partial x$ and $\mathcal{H} = L_2(0,1)$, the domain of L can be chosen as the set of all absolutely continuous functions on $0 \leq x \leq 1$, a set that is dense in $\mathcal{H}$ but smaller than $\mathcal{H}$.

If the problem (2.1-3) is well posed, the evolution operator is a bounded linear operator from $\mathcal{H}$ to $\mathcal{B}$. Since this evolution operator is bounded, its domain can be extended in a standard way from the domain of L to the whole space $\mathcal{H}$ (Richtmyer and Morton, 1967, p. 34). For notational convenience we shall assume henceforth that L is time independent so that the evolution operator is $\exp(Lt)$. In this case the formal solution of (2.1-3) is

$$u(t) = e^{Lt}u(0) + \int_0^t e^{L(t-s)}f(s)ds \qquad (2.4)$$

This formal solution is justified under the conditions that $f(t)$, $Lf(t)$, and $L^2f(t)$ exist and are continuous functions of t in the norm $||\cdot||$ for all $t \geq 0$ (see Richtmyer and Morton, 1967).

The semi-discrete approximations to (2.1) to be studied here are of the form

$$\frac{\partial u_N(x,t)}{\partial t} = L_N\, u_N(x,t) + f_N(x,t) \qquad (2.5)$$

where, for each t, $u_N(x,t)$ belongs to an N-dimensional subspace $\mathcal{B}_N$ of $\mathcal{B}$, and L_N is a linear operator from $\mathcal{H}$ to $\mathcal{B}_N$ of the form

$$L_N = P_N\, L\, P_N \quad . \qquad (2.6)$$

Here P_N is a projection operator of $\mathcal{H}$ onto $\mathcal{B}_N$ and $f_N = P_N f$. We shall assume that $\mathcal{B}_N \subset \mathcal{B}_M$ when $N < M$. For definiteness, we shall also assume the initial conditions for the approximate equations (2.5) to be $u_N(0) = P_N u(0)$ where $u(0) = g(x)$ is the initial condition (2.3). Specific examples of projections P_N and the resulting approximations L_N will be given below.

According to this general framework, the formulation of a spectral method involves two essential steps: (i) the choice of approximation space $\mathcal{B}_N$; and (ii) the choice of the projection operator P_N. It will turn out that the mathematical analysis of the methods also involves two key steps: (i) the analysis of how well functions in $\mathcal{H}$ can be approximated by functions in $\mathcal{B}_N$ (see Section 3) and, in particular, the estimation of $||u - P_N u||$ for arbitrary $u \in \mathcal{H}$; and (ii) the study of the 'stability' of L_N (see Section 4). Finally, there are the

important practical questions of how to time difference (see Section 10) and how to implement spectral methods efficiently (see Section 11). All these considerations will be tied together in Section 14 when we summarize our results on choosing a spectral method.

Galerkin or spectral approximation

A Galerkin approximation to (2.1-3) is constructed as follows (Collatz 1960, Orszag 1971a). The approximation u_N is sought in the form of the truncated series

$$u_N(x,t) = \sum_{n=1}^{N} a_n(t)\ \phi_n(x) \qquad (2.6)$$

where the time-independent functions ϕ_n are assumed linearly independent and $\phi_n \in \mathcal{B}_N$ for all n. Thus $u_N(x,t)$ necessarily satisfies all the boundary conditions. The expansion coefficients $a_n(t)$ are determined by the Galerkin equations

$$\frac{d}{dt}\ (\phi_n, u_N) = (\phi_n, L\ u_N) + (\phi_n, f) \qquad (n=1,\ldots,N) \qquad (2.7)$$

or

$$\sum_{m=1}^{N} (\phi_n, \phi_m)\ \frac{da_m}{dt} = \sum_{m=1}^{N} a_m(\phi_n, L\phi_m) + (\phi_n, f)\ .$$

These implicit equations for $a_n(t)$ can be put into the standard explicit form (2.4-5) by defining the projection operator P_N by

$$P_N u(x) = \sum_{n=1}^{N} \sum_{m=1}^{N} p_{nm}(\phi_m, u)\ \phi_n(x) \qquad (2.8)$$

where p_{nm} are the elements of the inverse of the $N \times N$ matrix whose elements are (ϕ_n, ϕ_m) .

Example 2.1: Fourier sine series

If we choose $H = L_2(0,\pi)$ and $\phi_n(x) = \sin nx$, we recover the Galerkin approximations given in Example 1.1-2 for the heat equation and in Example 1.3 for the wave equation. Every function $u \varepsilon L_2(0,\pi)$ has a Fourier sine series that converges in the L_2 norm, so that $||u - P_N u|| \to 0$ as $N \to \infty$. However, as illustrated by Example 1.3, this does not ensure that u_N converges to u as $N \to \infty$.

Example 2.2: Chebyshev series

We choose H to be the space of functions on the interval $|x| \leq 1$ that are square integrable with respect to the weight function $1/\sqrt{1-x^2}$. If the problem is

$$u_t + u_x = f(x,t) \quad (-1 \leq x \leq 1, \quad t > 0) , \tag{2.9a}$$

$$u(-1,t) = 0 , \quad u(x,0) = g(x) , \tag{2.9b}$$

which is a slight generalization of Example 1.3, it is appropriate to choose the expansion functions for the Galerkin approximates to be $\phi_n(x) = T_n(x) - (-1)^n T_0(x)$. Here $T_n(x)$ is the Chebyshev polynomial of degree n definied by $T_n(\cos\theta) = \cos n\theta$ when $x = \cos\theta$; thus, $T_0(x) = 1$, $T_1(x) = x$, $T_2(x) = 2x^2 - 1$, $T_3(x) = 4x^3 - 3x, \ldots,$. Observe that $\phi_n(x)$ satisfies the boundary condition

$\phi_n(-1) = 0$ because $T_n(-1) = (-1)^n$ for all n. The properties of Chebyshev polynomials are summarized in Section 15.

The Galerkin equations (2.7) are obtained explicitly as follows. First the the definition of $T_n(x)$ and the substitution $x = \cos\theta$ imply that

$$(T_n, T_m) = \int_0^\pi \cos n\theta \cos m\theta \, d\theta = \frac{\pi}{2} c_n \delta_{nm},$$

where

$$(f,g) = \int_{-1}^{1} f(x)g(x) \Big/ \sqrt{1-x^2} \, dx$$

Here $c_0 = 2$, $c_n = 1$ $(n > 0)$ and $\delta_{nm} = 0$ if $n \neq m$, 1 if $n = m$. Therefore,

$$(\phi_n, \phi_m) = \frac{\pi}{2} c_n \delta_{nm} + (-1)^{n+m} \pi .$$

Next, the Chebyshev polynomials satisfy

$$2T_n(x) = \frac{T'_{n+1}(x)}{n+1} - \frac{T'_{n-1}(x)}{n-1} \qquad (n \geq 2),$$

as may be verified by substituting $x = \cos\theta$. Therefore,

$$(\varphi_n, \phi'_m) = \begin{cases} \pi(-1)^{n+1} m + \pi m & n < m, \quad m + n \text{ odd} \\ \pi(-1)^{n+1} m & n > m, \quad m + n \text{ odd} \\ 0 & n + m \text{ even} \end{cases}$$

Using these results, (2.7) gives the Galerkin approximation

$$\frac{da_n}{dt} + 2(-1)^n \frac{d}{dt} \sum_{m=1}^{N} (-1)^m a_m = -2 \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{N} p\, a_p +$$

$$+ 2(-1)^n \sum_{\substack{p=1 \\ p \text{ odd}}}^{N} p\, a_p + \hat{f}_n + 2(-1)^n \hat{f}_0 \quad (n=1,\ldots,N)$$

Here $\hat{f}_n = (T_n, f)$ for $n = 0,\ldots,N$.

These Galerkin equations can be simplified by introducing the notation $a_0 = - \sum_{m=1}^{N} (-1)^m a_m$, so that (2.6) becomes

$$u_N(x,t) = \sum_{n=0}^{N} a_n(t) T_n(x) . \tag{2.10}$$

Substituting the above expression for a_0, the Galerkin dynamical equations can be rewritten as

$$\frac{da_n}{dt} = - \frac{2}{c_n} \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{N} p\, a_p + \hat{f}_n + \frac{1}{c_n} b(t) (-1)^n \quad (n=0,\ldots,N), \tag{2.11}$$

$$\sum_{n=0}^{N} (-1)^n a_n = 0 . \tag{2.12}$$

Here $b(t)$ is a 'boundary' term that ensures maintenance of the boundary condition (2.12). Using (2.12) it is easy to show that

$$b(t) = \frac{-1}{N+\frac{1}{2}} \left[\sum_{n=0}^{N} (-1)^n (n^2 a_n + \hat{f}_n) \right] = \frac{1}{N+\frac{1}{2}} \left[\frac{\partial u_N}{\partial x} \bigg|_{x=-1} - \sum_{n=0}^{N} (-1)^n \hat{f}_n \right]$$

Tau approximation

The tau approximation (Lanczos 1956) is obtained by choosing the expansion functions ϕ_n to be elements of a complete set of orthonormal function $\phi_n (n=1,2,...)$. The solution $u_N(x,t)$ is assumed expanded in the series

$$u_N(x,t) = \sum_{n=1}^{N+k} a_n(t)\phi_n(x) \ , \qquad (2.13)$$

which is similar to (2.6), but now the expansion functions ϕ_n are not required individually to satisfy the boundary constraints (2.2). Here k is the number of independent boundary constraints $Bu_N = 0$ that must be applied. The constraints

$$\sum_{n=1}^{N+k} a_n \, B \, \phi_n = 0 \qquad (2.14)$$

are imposed as part of the conditions determining the expansion coefficients a_n of a function in $\mathcal{B}_N$. The projection operator P_N is defined by

$$P_N\left(\sum_{n=1}^{\infty} A_n\phi_n\right) = \sum_{n=1}^{N} A_n\phi_n + \sum_{m=1}^{k} b_m\phi_{N+m} \qquad (2.15)$$

where b_m $(m=1,\ldots,k)$ are chosen so that the boundary constraints $Bu = 0$ are satisfied. It follows from these definitions that the tau approximation to (2.1-2) is given by (2.13) with the k equations (2.14) and the N equations

$$\frac{da_n}{dt} = (\phi_n, L\, u_N) + (\phi_n, f) \qquad (n=1,\ldots,N) \tag{2.16}$$

An equivalent formulation of the tau method is given as follows: The equations for the expansion coefficients a_n of the exact solution u in terms of the complete orthonormal basis ϕ_n are

$$u(x,t) = \sum_{n=1}^{\infty} a_n(t)\phi_n(x) \ ,$$

$$\frac{da_n}{dt} = (\phi_n, Lu) + (\phi_n, f) \qquad (n=1,2,\ldots) \tag{2.17}$$

The tau approximation equations for the $N+k$ expansion coefficients of u_N in (2.13) are obtained from the first N equations (2.17) with u replaced by u_N and the k boundary conditions (2.14). The origin of the name 'tau method' is that the resulting approximation u_N is the exact solution to the modified problem

$$\frac{\partial u_N}{\partial t} = L\, u_N + f + \sum_{p=1}^{\infty} \tau_p(t)\phi_{N+p}(x) \tag{2.18}$$

which lies in $\mathcal{B}_N$ for all $t > 0$. For each initial value problem and choice of orthonormal basis ϕ_n, there is a choice of τ-coefficients such that $u_N \in \mathcal{B}_N$, namely

$$\tau_p = (\phi_{N+p}, \frac{\partial u_N}{\partial t} - Lu_N - f) \quad \text{for} \quad p = 1,2,\cdots.$$

Example 2.3: Fourier sine series

For all of the applications given in Example 2.1, Galerkin and tau approximations based on $\phi_n = \sqrt{\frac{2}{\pi}} \sin nx$ are identical (except for the scaling factor $\sqrt{2/\pi}$) since the orthonormal expansion functions ϕ_n satisfy the boundary conditions.

Example 2.4: Chebyshev series

If we choose $\phi_{n+1}(x) = \frac{1}{\sqrt{c_n}} T_n(x)$ where $c_0 = 2$, $c_n = 1$ $(n > 0)$ and apply the tau method to the problem (2.9) the result can be recast into the form of equations (2.10-12) with $b(t) = 0$ and (2.11) only applied for $n = 0,1,\ldots,N-1$ instead of $n=0,1,\ldots,N$. Thus, the tau equations for the one-dimensional wave problem (2.9) are

$$\frac{da_n}{dt} = -\frac{2}{c_n} \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{N} p\, a_p + \hat{f}_n \qquad (0 \le n \le N-1) \qquad (2.19)$$

$$\sum_{n=0}^{N} (-1)^n a_n(t) = 0 \qquad (2.20)$$

Example 2.5: Laguerre series

Here we choose H to be the space of functions that are square integrable on $0 \leq x < \infty$ with respect to the weight function e^{-x}. We choose the expansion functions to be $\phi_n(x) = L_n(x)$ where $L_n(x)$ is the (normalized) Laguerre polynomial of degree n. $L_n(x)$ has the properties $(L_n, L_m) = \delta_{nm}$, $L_n(0) = 1$, and $L_n' - L_{n+1}' = L_n$ for all n,m.

Suppose we wish to solve

$$u_t + u_x = f(x,t) \qquad (0 \leq x < \infty, \quad t > 0) \tag{2.21a}$$

$$u(0,t) = 0 \qquad u(x,0) = g(x) \tag{2.21b}$$

by seeking an approximate solution of the form

$$u_N(x,t) = \sum_{n=0}^{N} a_n(t) L_n(x) \tag{2.22}$$

It is readily verified that the tau approximation (2.17) is

$$\frac{da_n}{dt} = \sum_{p=n+1}^{N} a_p + (L_n, f) \qquad (n=0,1,\ldots,N-1) \tag{2.23}$$

while the boundary condition is

$$\sum_{n=0}^{N} a_n = 0 \tag{2.24}$$

Similarly, the Laguerre-tau approximation to the heat equation problem

$$u_t = u_{xx} + f(x,t) \qquad (0 \leq x < \infty, \quad t > 0)$$

$$u(0,t) = 0 \qquad u(x,0) = g(x) \tag{2.25}$$

is given by (2.22), (2.24) and

$$\frac{da_n}{dt} = \sum_{p=n+1}^{N} (p-n-1)a_p + (L_n, f) \qquad (n=0,1,\ldots,N-1) \tag{2.26}$$

Collocation or pseudospectral approximation

The projection operator P_N for collocation [sometimes called the method of selected points (Lanczos 1956) or pseudospectral approximation (Orszag 1971a)] is defined as follows. Let $x_1, x_2, \ldots, x_N$ be N points interior to the domain D. These points are called the collocation points. Also let $\phi_n(x)$ $(n=1,\ldots,N)$ be a basis for the approximation space $\mathcal{B}_N$ and suppose that $\det \phi_n(x_m) \neq 0$. Then for each $u \in H$

$$P_N u = \sum_{n=1}^{N} a_n \phi_n(x) \tag{2.27}$$

where the expansion coefficients a_n are the solutions of the equations

$$\sum_{n=1}^{N} a_n \phi_n(x_i) = u(x_i) \quad (i=1,\ldots,N) \quad . \tag{2.28}$$

Thus, collocation is characterized by the conditions that $P_N u(x_i) = u(x_i)$ for $i = 1,\ldots,N$ and $P_N u \in \mathcal{B}_N$. Notice that the results of collocation depend on both the points x_n and the functions $\phi_n(x)$ for $n = 1,\ldots,N$.

Example 2.6: Fourier sine series

If we wish to solve the problems formulated in Examples 1.1-3 by collocation instead of Galerkin or tau methods we proceed as follows. The space $H = L_2(0,\pi)$ and we choose the expansion functions $\phi_n(x) = \sin nx$ $(n=1,\ldots,N)$ and the collocation points $x_j = \pi j/(N+1)$ $(j=1,\ldots,N)$. The collocation equations

$$\sum_{n=1}^{N} a_n \sin \frac{\pi jn}{N+1} = u(x_j) \quad (j=1,\ldots,N) \tag{2.29}$$

have the solution

$$a_n = \frac{2}{N+1} \sum_{j=1}^{N} u(x_j) \sin \frac{\pi jn}{N+1} \quad (n=1,\ldots,N) \tag{2.30}$$

This result follows from the relation

$$\sum_{n=1}^{N} \sin \frac{\pi jn}{N+1} \sin \frac{\pi kn}{N+1} = \frac{N+1}{2} \delta_{jk}$$

valid for $0 < j,k < N+1$. Thus,

$$P_N u = \sum_{n=1}^{N} a_n \sin nx \tag{2.31}$$

where a_n is given by (2.30).

It follows from (2.29-31) that

$$P_N L P_N u = \sum_{n=1}^{N} b_n \sin nx$$

where $b_n = - n^2 a_n$ $(n=1,\ldots,N)$ if $L = \partial^2/\partial x^2$, and

$$b_n = \frac{2}{N+1} \sum_{\substack{m=1 \\ m+n \text{ odd}}}^{N} \frac{m \sin \frac{\pi n}{N+1}}{\cos \frac{\pi m}{N+1} - \cos \frac{\pi n}{N+1}} a_m \qquad (n=1,\ldots,N)$$

if $L = \partial/\partial x$.

Example 2.7: Chebyshev collocation for the wave equation

Suppose we wish to solve the one-dimensional wave problem (2.9) using collocation. An appropriate basis for the approximation space B_N is the set of functions $\phi_n(x) = T_n(x) - (-1)^n T_0(x)$ $(n=1,\ldots,N)$ introduced in our discussion of Example 2 above. We choose the collocation points to be the extrema of the Chebyshev polynomial $T_N(x)$ satisfying $|x| \leq 1$. Since $T_N(\cos \theta) = \cos N\theta$ these extrema lie at $x_j = \cos \frac{\pi j}{N}$, $j = 0,\ldots,N-1$. The point $x = - 1$ is also an extremum of

$T_N(x)$ but it is <u>not</u> included in the set of collocation points because the boundary conditions for (2.9) are imposed at $x = -1$ so $\phi_n(-1) = 0$ for all n.

As in Example 2.2, the expansion coefficients a_n for $n = 1,\ldots,N$ may be augmented by defining $a_0 = -\sum_{m=1}^{N} (-1)^m a_m$ so that

$$u_N(x,t) = \sum_{n=0}^{N} a_n(t) T_n(x) \ .$$

It may easily be shown that the collocation equations for $a_n(t)$ that follow from (2.9) are

$$\frac{da_n}{dt} = -\frac{2}{c_n} \sum_{\substack{p=n+1 \\ p+n \text{ odd}}}^{N} p\, a_p + f_n + \frac{1}{\bar{c}_n} b(t)(-1)^n \quad (n=0,\ldots,N) \qquad (2.32)$$

$$\sum_{n=0}^{N} (-1)^n a_n(t) = 0 \qquad (2.33)$$

where $f_n = (T_n, f)$ and $\bar{c}_0 = \bar{c}_N = 2$, $\bar{c}_n = 1$ $(0 < n < N)$. Here $b(t)$ is a 'boundary' term that is used to ensure compliance with the boundary condition (2.33). It is easy to show that

$$b(t) = -\frac{1}{N} \sum_{n=0}^{N} (-1)^n (n^2 a_n + \hat{f}_n) = \frac{1}{N} \left[\frac{\partial u_N}{\partial x}\Big|_{x=-1} - \sum_{n=0}^{N} (-1)^n \hat{f}_n \right]$$

The reader should observe the close similarity between the Chebyshev Galerkin, tau, and collocation equations for the problem (2.9). The only difference between them is the way the boundary term $b(t)$ enters. In the Galerkin equations (2.11), $b(t)$ appears with the coefficient $(-1)^n/c_n$; in the tau equations $b(t)$ enters with the coefficient δ_{nN} so it appears only in the equation for a_N as a tau coefficient; with collocation, the coefficient of $b(t)$ is $(-1)^n/\bar{c}_n$. This close similarity between the three methods for the wave equation can also be seen by observing that when $f(x,t)$ is a polynomial of degree N in x, all three approximation methods give Nth degree polynomial approximations $u_N(x,t)$ that satisfy <u>exactly</u> the initial value problem

$$\frac{\partial u_N}{\partial t} + \frac{\partial u_N}{\partial x} = f(x,t) + \tau(t)Q_N(x) \tag{2.34}$$

$$u_N(0,t) = 0 \ .$$

In the tau method, $Q_N(x) = T_N(x)$; in collocation,

$$Q_N(x) = \prod_{j=0}^{N-1} (x-x_j) = 2^{2-N} \sum_{n=0}^{N}{}'' \frac{(-1)^{n+N}}{\bar{c}_n} T_n(x) = \frac{1}{N} 2^{1-N}(x-1)T_N'(x)$$

where $x_j = \cos \frac{\pi j}{N}$ $(j = 0,\ldots,N-1)$ are the collocation points; finally, the Galerkin equations (2.10) are obtained if

$$Q_N(x) = \sum_{n=0}^{N}{}'' \frac{(-1)^n}{c_n} T_n(x) \ .$$

For all three methods $\tau(t)$ is uniquely determined by the requirement that $u_N(x,t)$ be a polynomial of degree N in x for all t.

Example 2.8: Chebyshev spectral methods for the heat equation

To illustrate further the nature of the differences between Galerkin, tau and collocation methods, we apply them to the heat equation

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} + f(x,t) \qquad (-1 < x < 1\,, \quad t > 0)$$

$$u(-1,t) = u(1,t) = 0 \quad (t > 0)\,, \qquad u(x,0) = g(x) \quad (-1 < x < 1).$$

We approximate $u(x,t)$ by

$$u_N(x,t) = \sum_{n=0}^{N} a_n(t) T_n(x)\,.$$

The Galerkin, tau, and collocation equations for $a_n(t)$ are all of the form

$$\frac{da_n}{dt} = \frac{1}{c_n} \sum_{\substack{p=n+2 \\ p+n \text{ even}}}^{N} p(p^2-n^2)a_p + \hat{f}_n(t) + b_1(t)B_{1n} + b_2(t)B_{2n} \qquad (2.35)$$

$$\sum_{n=0}^{N} a_n = \sum_{n=0}^{N} (-1)^n a_n = 0\,, \qquad (2.36)$$

where $\hat{f}_n = (T_n, f)$. Eqs. (2.36) are just a restatement of $u_N(\pm 1, t) = 0$. The terms $b_1(t)$ and $b_2(t)$ in (2.35) are boundary terms that ensure compliance with the boundary conditions (2.36). The only differences between the three approximation methods lies in the coefficients B_{1n} and B_{2n}.

In the tau method,

$$B_{1n} = \delta_{n,N-1} \, , \quad B_{2n} = \delta_{nN} \, . \tag{2.37a}$$

In the Galerkin method,

$$B_{1n} = \frac{1}{c_n} \quad , \quad B_{2n} = \frac{(-1)^n}{c_n} \, ; \tag{2.37b}$$

this result follows using the expansion functions

$$\phi_n(x) = T_n(x) - \begin{cases} T_0(x) & n \text{ even} \\ T_1(x) & n \text{ odd} \end{cases}$$

that satisfy $\phi_n(\pm 1) = 0$ and augmenting the expansion coefficients a_n for $n \geq 2$ by $a_0 = -\sum a_{2n}$ and $a_1 = -\sum a_{2n+1}$. Finally, with collocation performed at the points $x_j = \cos \frac{\pi j}{N}$ $(j = 1,2,\ldots,N-1)$ the coefficients B_{1n} and B_{2n} in (2.35) are given by

$$B_{1n} = \frac{1}{\bar{c}_n} \qquad B_{2n} = \frac{(-1)^n}{\bar{c}_n} \, . \tag{2.37c}$$

It may also be verified that the boundary terms $b_1(t)$ and $b_2(t)$ are of the form

$$b_i(t) = C_{i+} \left.\frac{\partial^2 u}{\partial x^2}\right|_{x=+1} + \sum_{n=0}^{N} f_n + C_{i-} \left.\frac{\partial^2 u}{\partial x^2}\right|_{x=-1} + \sum_{n=0}^{N} (-1)^n \hat{f}_n \qquad (2.38)$$

for $i = 1,2$. Here

$$C_{1+} = -\tfrac{1}{2} \quad , \quad C_{1-} = \tfrac{1}{2}(-1)^N \quad ,$$

$$C_{2+} = -\tfrac{1}{2} \quad , \quad C_{2-} = \tfrac{1}{2}(-1)^{N+1} \quad ,$$

for the tau method;

$$C_{1+} = -\frac{N+\frac{1}{2}}{N^2+N} \quad , \quad C_{1-} = \tfrac{1}{2}\frac{(-1)^N}{N^2+N} \quad ,$$

$$C_{2+} = \tfrac{1}{2}\frac{(-1)^N}{N^2+N} \quad , \quad C_{2-} = -\frac{N+\frac{1}{2}}{N^2+N} \quad ,$$

for the Galerkin method;

$$C_{1+} = -\frac{1}{N} \qquad C_{1-} = 0$$

$$C_{2+} = 0 \qquad C_{2-} = -\frac{1}{N}$$

for the collocation method.

In the previous examples the only difference between Galerkin, tau, and collocation approximations is their treatment of the boundary terms. However, in more complicated problems, there are significant differences between these approximations. The next example illustrates the influence of quadratic nonlinearity.

Example 2.9: Chebyshev approximations to Burgers' equation

Chebyshev series approximations to the solution $u(x,t)$ to Burgers' equation

$$\frac{\partial u}{\partial t} + u\frac{\partial u}{\partial x} = \nu\frac{\partial^2 u}{\partial x^2} \qquad (|x|\le 1, t>0) \tag{2.39}$$

$$u(\pm 1,t) = 0$$
$$u(x,0) = f(x)$$

are obtained by methods very similar to those for linear equations. In general, spectral approximations to the nonlinear equation

$$\frac{\partial u}{\partial t} = A(u) \tag{2.40}$$

are of the form

$$\frac{\partial u_N}{\partial t} = P_N A(P_N u_N) \tag{2.41}$$

where P_N is a projection operator. The projection operator P_N can be that for Galerkin, tau, or collocation approximations.

If we write

$$u_N(x,t) = \sum_{n=0}^{N} a_n(t)\, T_n(x) ,$$

then the Galerkin approximation to (2.39) is given by

$$c_n \frac{da_n}{dt} = -2 \sum_{\substack{|m|\le N \\ |p|\le N \\ m+p\ge n+1 \\ n+m+p \text{ odd}}} p\, \bar{a}_m \bar{a}_p + \nu \sum_{\substack{m=n+2 \\ m+n \text{ even}}} m(m^2-n^2) a_m + b_+(t) + b_-(t)(-1)^n \qquad (0\le n\le N), \tag{2.42a}$$

$$\sum_{n=0}^{N} a_n = \sum_{n=0}^{N} a_n (-1)^n = 0 \ , \qquad (2.42b)$$

where $\bar{a}_m = c_{|m|}\, a_{|m|}$ for $|m| \le N$. The tau equations are identical except that (2.42a) only applies for $0 \le n \le N-2$ and $b_+ = b_- = 0$. On the other hand, the collocation equations obtained using the collocation points $x_j = \cos \frac{\pi j}{N}$ for $j = 1,\ldots,N-1$ are just (2.42b) and

$$\bar{c}_n \frac{da_n}{dt} = -2 \sum_{\substack{|m|\le N \\ |p|\le N \\ m+p\ge n+1 \\ n+m+p \text{ odd}}} p\, \bar{a}_m \bar{a}_p \; -2 \sum_{\substack{|m|\le N \\ |p|\le N \\ m+p\ge 2N-n+1 \\ n+m+p \text{ odd}}} p\, \bar{a}_m \bar{a}_p$$

$$+ \nu \sum_{\substack{m=n+2 \\ m+n \text{ even}}}^{N} m(m^2-n^2) a_m + \bar{b}_+(t) + \bar{b}_-(t)(-1)^n \qquad (2.43)$$

$$(0 \le n \le N)$$

where $\bar{c}_0 = \bar{c}_N = 2$ and $\bar{c}_n = 1$ for $n \neq 0,N$. Observe the appearance of the 'aliasing' term as the second sum on the right side of (2.43). We shall discuss this term in more detail in Section 11.

Example 2.10: Chebyshev approximations to $u_t + F(u)_x = 0$

Galerkin and tau approximations to the solution to

$$u_t + F(u)_x = 0 \qquad (2.44)$$

where $F(u)$ is arbitrarily nonlinear, are very unwieldy both to write down explicitly and to solve on a computer. On the one hand,

collocation equations may also be hard to write down explicitly, they lend themselves to ready solution without their explict form being known!

The collocation approximation to (2.44) is obtained as follows. We use the relation

$$(F(u_N))_x = F'(u_N)\ \frac{\partial u_N}{\partial x}\ . \tag{2.45}$$

Since $\partial u_N/\partial x$ can be computed explicitly in terms of u_N as a polynomial in x of degree $N-1$, it follows that $(F(u_N))_x$ can be evaluated by this formula at each of the collocation points assuming that $F'(z)$ is a known function; thus, the collocation approximation to (2.44) is determined.

There is a slightly different collocation procedure that can also be applied to (2.44). It has the operator form

$$\frac{\partial u_N}{\partial t} + P_N\ \frac{\partial}{\partial x}\ P_N\ F(u_N) = 0 \tag{2.46}$$

which is usually not the same as the collocation approximation of the form (2.41) described above. However, since $P_N F(u_N)$ can be computed by collocation from u_N and since the collocation approximation to $P_N \partial/\partial x$ has already been given in Example 2.7, $\partial u_N/\partial t$ is determined by (2.46). The collocation approximation given by (2.41) or (2.45) differs from (2.46) by the term

$$P_N\ \frac{\partial}{\partial x}\ (I-P_N)\ F(u_N)$$

which is generally not zero. However, if $F'(z)$ is not known accurately then (2.46) may be the only viable method. More details on these collocation algorithms are given in Section 11.

3. Survey of Approximation Theory

The remarkable convergence properties of spectral methods to be discussed later owe to the rapid convergence of expansions of smooth functions in series of orthogonal functions. We present a summary of the relevant theory here.

Fourier series

The complex Fourier series of $f(x)$ defined for $0 \le x \le 2\pi$ is the periodic function

$$g(x) = \sum_{k=-\infty}^{\infty} a_k e^{ikx} \tag{3.1}$$

where

$$a_k = \frac{1}{2\pi} \int_0^{2\pi} f(x) e^{-ikx} dx \ . \tag{3.2}$$

We shall show below that if $f(x)$ is piecewise continuous and has bounded total variation then

$$g(x) = \tfrac{1}{2}[f(x+)+f(x-)] \tag{3.3}$$

for $0 \le x \le 2\pi$ and $g(x)$ is repeated periodically outside $0 \le x \le 2\pi$. In particular, $g(0) = g(2\pi) = \tfrac{1}{2}[f(0+)+f(2\pi-)]$.

The Fourier sine series of a function $f(x)$ defined for $0 < x < \pi$ is the function

$$g_s(x) = \sum_{k=1}^{\infty} a_k \sin kx \tag{3.4}$$

where

$$a_k = \frac{2}{\pi} \int_0^{\pi} f(x) \sin kx \, dx \, . \tag{3.5}$$

The Fourier cosine series of a function defined for $0 < x < \pi$ is

$$g_c(x) = \sum_{k=0}^{\infty} a_k \cos kx \tag{3.6}$$

where

$$a_k = \frac{2}{\pi c_k} \int_0^{\pi} f(x) \cos kx \, dx \tag{3.7}$$

with $c_0 = 2$, $c_k = 1$ $(k > 0)$. It follows easily from (3.3) that if $f(x)$ is piecewise continuous and of bounded total variation then

$$g_s(x) = f_s(x) \tag{3.8}$$

$$g_c(x) = f_c(x) \tag{3.9}$$

where $f_s(x) = f_c(x) = \frac{1}{2}[f(x+)+f(x-)]$ for $0 < x < \pi$, $f_s(-x) = -f_s(x)$, $f_c(-x) = f_c(x)$ for $-\pi < x < 0$, $f_s(0) = f_s(\pi) = 0$, $f_c(0) = f(0+)$, $f_c(\pi) = f(\pi-)$, and $f_s(x)$ and $f_c(x)$ are extended periodically outside the interval $-\pi < x \leq \pi$.

Convergence of Fourier series

To prove (3.3) we define $g_K(x)$ as the partial sum

$$g_K(x) = \sum_{k=-K}^{K} a_k e^{ikx} \tag{3.11}$$

Using (3.2) and the trigonometric sum formula

$$\sum_{k=-K}^{K} e^{iks} = \frac{\sin[(K+\frac{1}{2})s]}{\sin(\frac{1}{2}s)} \quad ,$$

we obtain

$$g_K(x) = \frac{1}{2\pi} \int_{x-2\pi}^{x} \frac{\sin[(K+\frac{1}{2})t]}{\sin[\frac{1}{2}t]} f(x-t)dt \qquad (3.12)$$

The kernel $\sin(K+\frac{1}{2})t/\sin\frac{1}{2}t$ of the integral (3.2) is plotted for several values of K in Figure 3.1. This figure suggests that when $f(x)$ has bounded total variation the leading contribution to the integral as $K \to \infty$ comes from the neighborhood of $t = 0$ since the contributions from the rest of the integration region should nearly cancel due to the rapid oscillations of the integrand. Thus,

$$g_K(x) \sim \frac{1}{2\pi} \int_{-\varepsilon}^{+\varepsilon} \frac{\sin[(K+\frac{1}{2})t]}{\sin[\frac{1}{2}t]} f(x-t)dt \qquad (K\to\infty) \qquad (3.13)$$

for any fixed $\varepsilon > 0$. Since ε may be chosen small we may replace $\sin\frac{1}{2}t$ by $\frac{1}{2}t$ with a maximum error of $0(\varepsilon^3)$. Also since $f(x-t)$ is piecewise continuous, we may assume that $f(x-t)$ is continuous for $0 \leq t \leq \varepsilon$ and $-\varepsilon \leq t < 0$ with at worst a jump discontinuity at $t = 0$. Therefore we may replace $f(x-t)$ by $f(x-)$ for $t > 0$ and $f(x+)$ for $t < 0$ giving

$$g_K(x) \sim [f(x+)+f(x-)]\,\frac{1}{\pi}\int_0^{\varepsilon}\frac{\sin(K+\frac{1}{2})s}{s}\,ds \qquad (K\to\infty)$$

Since

$$\frac{1}{\pi}\int_0^{\varepsilon}\frac{\sin(K+\frac{1}{2})s}{s}\,ds = \frac{1}{\pi}\int_0^{(K+\frac{1}{2})\varepsilon}\frac{\sin s}{s}\,ds \sim \frac{1}{\pi}\int_0^{\infty}\frac{\sin s}{s}\,ds = \tfrac{1}{2} \quad (K\to\infty)$$

for any fixed $\varepsilon > 0$, we find that

$$g_K(x) \sim \tfrac{1}{2}[f(x+)+f(x-)] \qquad (K\to\infty)$$

proving (3.3).

In the neighborhood of a point of discontinuity of $f(x)$ [or $x = 0$ and $x = 2\pi$ if $f(0+) \neq f(2\pi-)$] the convergence of $g_K(x)$ to its limit (3.3) as $K \to \infty$ is not uniform. To investigate the detailed approach of $g_K(x)$ to $g(x)$ near a point of discontinuity x_0 of $f(x)$, we use the asymptotic integral representation (3.13) to obtain

$$g_K(x_0+\frac{z}{K+\frac{1}{2}}) \sim \frac{1}{\pi}\int_{-\varepsilon}^{\varepsilon}\frac{\sin[(K+\frac{1}{2})t]}{t}\,f(x_0+\frac{z}{K+\frac{1}{2}}-t)\,dt \qquad (K\to\infty)$$

for every fixed z. Since ε is assumed small we can approximate $f(x_0+s)$ by $f(x_0+)$ for $0 < s < \varepsilon$ and by $f(x_0-)$ for $-\varepsilon < s < 0$. Therefore, for each fixed z and ε,

$$g_K(x_0+\frac{z}{K+\frac{1}{2}}) \sim \frac{f(x_0+)}{\pi}\int_{-\varepsilon}^{z/(K+\frac{1}{2})}\frac{\sin(K+\frac{1}{2})}{t}\,dt + \frac{f(x_0-)}{\pi}\int_{z/(K+\frac{1}{2})}^{\varepsilon}\frac{\sin(K+\frac{1}{2})t}{t} \qquad (K\to\infty)$$

$$= \frac{f(x_0+)}{\pi}\int_{-\varepsilon(K+\frac{1}{2})}^{z}\frac{\sin s}{s}\,ds + \frac{f(x_0-)}{\pi}\int_{z}^{\varepsilon(K+\frac{1}{2})}\frac{\sin s}{s}\,dx \qquad (K\to\infty)$$

$$\sim \frac{f(x_0+)}{\pi}\int_{-\infty}^{z}\frac{\sin s}{s}\,dx + \frac{f(x_0-)}{\pi}\int_{z}^{\infty}\frac{\sin s}{s}\,ds \qquad (K\to\infty)$$

Since $\int_{-\infty}^{\infty} \sin s/s\, ds = \pi$, we obtain

$$g_K(x_0+\frac{z}{K+\frac{1}{2}}) \sim \tfrac{1}{2}[f(x_0+)+f(x_0-)] + \frac{1}{\pi}[f(x_0+)-f(x_0-)]\,\mathrm{Si}(z) \qquad (K\to\infty)$$

(3.14a)

for any fixed z. Here the sine integral $\mathrm{Si}(z)$ is defined

$$\mathrm{Si}(z) = \frac{1}{\pi}\int_0^z \frac{\sin s}{s}\,ds \qquad (3.14b)$$

A plot of $\frac{1}{\pi}\mathrm{Si}(z)$ is given in Figure 3.2.

The result (3.14) shows that if $x - x_0 = 0(\frac{1}{K})$ as $K \to \infty$ then $g_K(x) - \frac{1}{2}[f(x_0+)+f(x_0-)] = 0(1)$. This shows the nonuniformity of convergence of $g_K(x)$ to $f(x)$ in the neighborhood of the discontinuity x_0. The detailed description of this nonuniform limit given by (3.14) is called the Gibbs phenomenon. To illustrate the Gibbs phenomenon in an actual Fourier series, we plot in Figure 3.3 the partial sums of the Fourier sine series expansion of the function

$$f(x) = x/\pi \qquad (0<x<\pi)$$

The extended function $f_S(x)$ is discontinuous at $x = \pi$ leading to the Gibbs phenomenon there.

As $K \to \infty$, the maximum error of the partial sums of a Fourier (complex or sine or cosine) series in the neighborhood of a point of discontinuity occurs at the maximum of $Si(z)$ given by (3.14b). Since $Si'(z) = 0$ when $z = n\pi$ for $n = \pm 1, \pm 2, \ldots,$ the maximum error must occur at one of these points. It is easy to argue that the maximum of $Si(z)$ actually occurs at $z = \pi$ where

$$\frac{1}{\pi} Si(\pi) \doteq .58949 \tag{3.15}$$

Thus the maximum overshoot of the partial sums of the Fourier series near a discontinuity occurs near $x = x_0 + \frac{\pi}{K+\frac{1}{2}}$ for K large and is of magnitude

$$g_K(x_0+\frac{\pi}{K+\frac{1}{2}}) - f(x_0+) \sim .08949[f(x_0+)-f(x_0-)] \qquad (K\to\infty) \tag{3.16}$$

where the quantity in square brackets is the jump at x_0. For the example ploted in Figure 3.3 the jump of $f_S(x)$ at $x = \pi$ has magnitude 2 so the Fourier series gives a local overshoot of magnitude 0.179.

As $z \to \pm \infty$, $Si(z) \to \pm \frac{1}{2}\pi$ so that (3.14) is consistent with the convergence of the Fourier series to $f(x_0+)$ just to the right of x_0 and to $f(x_0-)$ just to the left of x_0. The Gibbs phenomenon only appears when $x \to x_0$ at the rate $1/K$ as $K \to \infty$.

Rate of Convergence of Fourier Series

If $f(x)$ is smooth and periodic, its Fourier series does not exhibit the Gibbs phenomenon. The Fourier series of $f(x)$ converges rapidly and uniformly. If $f(x)$ has continuous derivatives

of order $p = 0,1,\ldots,n-1$ and $f^{(n)}(x)$ is integrable, then by applying integration by parts to (3.2) and recalling the Riemann-Lebesgue lemma, we obtain

$$a_k << 1/k^n \qquad (k \to \pm\infty) \tag{3.17}$$

Here continuity of $f^{(p)}(x)$ also requires $f^{(p)}(0) = f^{(p)}(2\pi)$. For example, if $f(x)$ is continuous with $f(0) = f(2\pi)$ and $f'(x)$ is integrable then $a_k << 1/k$ as $k \to \infty$; if, in addition, $f'(x)$ is piecewise continuous and f'' is integrable then $a_k = 0(1/k^2)$ as $k \to \infty$.

Now we can be more precise in our estimates of the error $g_K(x) - f(x)$. If a_k goes to zero like $1/k^n$ as $k \to \infty$ then $f^{(n-1)}(x)$ is discontinuous. In this case,

$$g_K(x) - f(x) = 0\left(\frac{1}{K^n}\right) \qquad (K \to \infty) \tag{3.18}$$

when x is fixed away from a point of discontinuity of $f^{(n-1)}$ as $K \to \infty$, while

$$g_K(x) - f(x) = 0\left(\frac{1}{K^{n-1}}\right) \qquad (K \to \infty) \tag{3.19}$$

when $x - x_0 = 0\left(\frac{1}{K}\right)$ as $K \to \infty$ where x_0 is a point of discontinuity of $f^{(n-1)}(x)$.

In particular, if $f(x)$ is infinitely differentiable and periodic $[f(x+2\pi) = f(x)]$, $g_K(x)$ converges to $f(x)$ more rapidly than any finite power of $1/K$ as $K \to \infty$ for all x .

Fourier sine and cosine series have convergence properties very similar to the complex Fourier series just discussed. We summarize these properties for Fourier cosine series. If derivatives of $f(x)$ of order $p = 0,1,\ldots,n-1$ are continuous for $0 < x < \pi$ while $f^{(p)}(0) = f^{(p)}(\ ) = 0$ for all <u>odd</u> $p < n$ and $f^{(n)}(x)$ is integrable, then the Fourier cosine coefficients given by (3.7) satisfy

$$a_n << 1/k^n \qquad (k \to \infty) \tag{3.20}$$

as may be proven by integration by parts.

Thus, if $f(x)$ is infinitely differentiable for $0 \leq x \leq \pi$ and $f^{(2p+1)}(0) = f^{(2p+1)}(\pi) = 0$ for $p = 0,1,\ldots$ then the Fourier cosine coefficients a_k approach zero more rapidly than any power of $1/k$ as $k \to +\infty$. In other words, if $f(x)$ is infinitely differentiable on $-\infty \leq x \leq \infty$, periodic with period 2π $[f(x+2\pi) = f(x)]$, and even $[f(x) = f(-x)]$, then the remainder after N terms of the Fourier cosine series (3.6) goes to zero more rapidly than any finite power of $1/N$ as $K \to \infty$.

To compare the convergence properties of Fourier sine and cosine series, we have plotted in Figures 3.3 and 3.4 some results for the Fourier sine and cosine expansions, respectively, of the function x/π for $0 \leq x \leq \pi$. As discussed above, the Gibbs phenomenon in the sine series expansion is evident at $x = \pi$ (see Figure 3.3). Observe that the error in the N term partial sum goes to zero like $1/N$ as $N \to \infty$ when x is fixed $0 \leq x < \pi$. In Figure 3.4, we plot the error between the N term cosine series and x/π. Observe that as $N \to \infty$ the error goes to zero like $1/N^2$ for $0 < x < \pi$ and like $1/N$ when $x = 0(1/N)$ as $N \to \infty$.

Chebyshev polynomial expansions

The convergence theory of Chebyshev polynomial expansions is very similar to that of Fourier cosine series. In fact, if

$$g(x) = \sum_{k=0}^{\infty} a_k T_k(x) \tag{3.21}$$

is the Chebyshev series associated with $f(x)$ for $-1 \leq x \leq 1$ then $G(\theta) = g(\cos\theta)$ is the Fourier cosine series of $F(\theta) = f(\cos\theta)$ for $0 \leq \theta \leq \pi$. This result follows from the definition of $T_n(x)$: since $T_n(\cos\theta) = \cos n\theta$,

$G(\theta) = g(\cos\theta) = \sum_{k=0}^{\infty} a_n \cos n\theta$. Thus,

$$a_k = \frac{2}{\pi c_k} \int_0^{\pi} f(\cos\theta)\cos k\theta \, d\theta = \frac{2}{\pi c_k} \int_{-1}^{1} f(x)T_k(x)(1-x^2)^{-\frac{1}{2}} \, dx \tag{3.22}$$

where $c_0 = 2$, $c_k = 1$ $(k > 0)$.

It follows from this close relation between Chebyshev series and Fourier cosine series that if $f(x)$ is piecewise continuous and if $f(x)$ is of bounded total variation for $-1 \leq x \leq 1$ then $g(x) = \frac{1}{2}[f(x+)+f(x-)]$ for each x $(-1 < x < 1)$ and $g(1) = f(1-)$, $g(-1) = f(-1+)$. Also, if $f^{(p)}(x)$ is continuous for all $|x| \leq 1$ for $p = 0,1,\ldots,n-1$, and $f^{(n)}(x)$ is integrable, then

$$a_k << 1/k^n \qquad (k \to \infty). \tag{3.23}$$

Since $|T_k(x)| \leq 1$ for $|x| \leq 1$, it follows that the remainder after K terms of the Chebyshev series (3.23) is very much smaller than $1/K^{n-1}$ as $K \to \infty$. If $f(x)$ is infinitely differentiable for $|x| \leq 1$, the error in the Chebyshev series goes to zero more rapidly than any finite power of $1/K$ as $K \to \infty$.

The most important feature of Chebyshev series is that their convergence properties are not affected by the values of $f(x)$ or its derivatives at the boundaries $x = \pm 1$ but only by the smoothness of $f(x)$ and its derivatives throughout $-1 \leq x \leq 1$. In contrast, the Gibbs phenomenon shows that the rate of convergence of Fourier series depends on the values of f and its derivatives at the boundaries in addition to the smoothness of f in the interior of the interval. The reason for the absence of a Gibbs phenomenon for the Chebyshev series of $f(x)$ and its derivatives at $x = \pm 1$ is due to the fact that $F(\theta) = f(\cos\theta)$ satisfies $F^{(2p+1)}(0) = F^{(2p+1)}(\pi) = 0$ provided only that all derivatives of $f(x)$ of order at most p exist at $x = \pm 1$.

While Chebyshev expansions do not exhibit the Gibbs phenomenon at the boundaries $x = \pm 1$, they do exhibit the phenomenon at any interior discontinuity of $f(x)$. In Figure 3.5 we plot the partial sums of the Chebyshev expansions of the sign function sgn x:

$$\operatorname{sgn} x = \frac{4}{\pi} \sum_{n=0}^{\infty} (-1)^n \frac{T_{2n+1}(x)}{2n+1} \tag{3.24}$$

Near $x = 0$, a Gibbs phenomenon is observed while for $x \neq 0$ the error after N terms is of order $1/N$. In general, the local

structure of the partial sums $g_K(x)$ of Chebyshev series near a discontinuity of $f(x)$ is, aside from a simple scaling given by (3.14):

$$g_K\left(x_0+\frac{z}{K+\frac{1}{2}\sqrt{1-x_0^2}}\right) \sim \frac{1}{2}[f(x_0+)+f(x_0-)] + \frac{1}{\pi}[f(x_0+)-f(x_0-)]\ \mathrm{Si}(z) \qquad (K\to\infty)$$

where $|x_0| < 1$ and z is fixed. This equation is derived by a simple extension of the argument used to derive (3.14) [cf. (3.33) below for the explanation of the origin of the scale factor $1/\sqrt{1-x_0^2}$].

Rate of convergence of Sturm-Liouville eigenfunction expansions

Let us consider the expansion of a function $f(x)$ in terms of the eigenfunctions ϕ_n of a Sturm-Liouville problem: The eigenfunction $\phi_n(x)$ is a nonzero solution to

$$\frac{d}{dx}\,p(x)\frac{d\phi_n}{dx} + (\lambda_n w(x)-q(x))\ \phi_n(x) = 0 \qquad (3.25)$$

satisfying homogeneous boundary conditions. To be specific in our discussion we assume the boundary conditons $\phi_n(a) = \phi_n(b) = 0$, although the analysis applies more generally. We assume that $p(x) \geq 0$, $w(x) \geq 0$, $q(x) \geq 0$ for $a \leq x \leq b$. We will also assume that the eigenfunctions are normalized so that they satisfy

$$\int_a^b w(x)\phi_n(x)\phi_m(x)dx = \delta_{nm} \quad . \qquad (3.26)$$

and that they form a complete set; the latter property follows if $\lambda_n \to \infty$ as $n \to \infty$ (see Courant & Hilbert, 1953, p. 424).

We wish to estimate the rate of convergence of the eigenfunction expansion

$$f(x) = \sum_{n=1}^{\infty} a_n \phi_n(x) \quad . \tag{3.27a}$$

Using the orthonormality relation (3.26), the L_2 - error after N terms is

$$\int_a^b \left| f(x) - \sum_{n=1}^{N} a_n \phi_n(x) \right|^2 w(x)dx = \sum_{n=N+1}^{\infty} a_n^2 \quad ,$$

which may be estimated by calculating the rate of decrease of a_n as $n \to \infty$.

Orthonormality of $\{\phi_n\}$ implies that

$$a_n = \int_a^b f(x)\phi_n(x)w(x)dx \quad . \tag{3.27b}$$

Substituting $w(x)\phi_n(x)$ from the Sturm-Liouville equation (3.25) gives

$$a_n = \frac{1}{\lambda_n} \int_a^b \left(- \frac{d}{dx} p(x) \frac{d\phi_n}{dx} + q(x)\phi_n \right) f(x)dx \quad .$$

Integrating twice by parts, we obtain

$$a_n = \frac{1}{\lambda_n} p(x)[\phi_n(x)f'(x) - \phi_n'(x)f(x)]\Big|_{x=a}^{b} + \frac{1}{\lambda_n} \int_a^b h(x)\phi_n(x)w(x)dx$$

(3.28)

where

$$h(x) = \left[-\frac{d}{dx} p(x) \frac{df}{dx} + q(x) f(x) \right] / w(x). \qquad (3.29)$$

This integration by parts is justified if f is twice differentiable and h is square integrable with respect to w. Under these conditions and recalling the $\phi_n(a) = \phi_n(b) = 0$, we obtain

$$a_n = \frac{1}{\lambda_n} [p(a)\phi_n'(a)f(a) - p(b)\phi_n'(b)f(b)] + 0(\frac{1}{\lambda_n})$$

as $n \to \infty$, since $\left| \int_a^b h\phi_n w \, dx \right|^2 \leq \int_a^b h^2 w \, dx \int_a^b \phi_n^2 w \, dx = 0(1)$ as $n \to \infty$.

Nonsingular Sturm-Liouville problems

To proceed further we must distinguish between nonsingular and singular Sturm-Liouville problems: a problem is nonsingular if $p(x) > 0$ and $w(x) > 0$ throughout $a \leq x \leq b$. The important conclusion from (3.29) is that if the Sturm-Liouville problem is nonsingular and if $f(a)$ or $f(b)$ is nonzero then

$$a_n \sim \frac{1}{\lambda_n} [p(a)\phi_n'(a)f(a) - p(b)\phi_n'(b)f(b)] \quad (n \to \infty) \qquad (3.30)$$

(Notice that if $\phi_n'(a) = 0$, then $\phi_n(x) \equiv 0$ since (3.25) is second-order and $p(x) \neq 0$). It is well known [Courant & Hilbert 1953] that the asymptotic behavior of the eigenvalues and eigenfunctions of a nonsingular Sturm-Liouville problem are given by

$$\lambda_n \sim \left[n\pi / \int_a^b \sqrt{\frac{w}{p}}\, dx\right]^2 \qquad (n \to \infty) \qquad (3.31)$$

$$\phi_n(x) \sim A_n \sin\left(\sqrt{\lambda_n} \int_a^x \sqrt{\frac{w}{p}}\, dx\right) \qquad (n \to \infty) \qquad (3.32)$$

Using these relations in (3.30), we find that a_n behaves like $\frac{1}{n}$ as $n \to \infty$. This behavior of a_n leads to the Gibbs phenomenon in the expansion (3.26) near the boundary points at which $f(a)$ or $f(b) \neq 0$. The asymptotic behavior (3.31-32) ensures that this Gibbs phenomenon is asymptotically the same as that for Fourier sine series in terms of the stretched independent variable

$$X = \pi(x-a)\sqrt{w(a)/p(a)} \Big/ \int_a^b \sqrt{w(x)/p(s)}\, ds \qquad (3.33)$$

near $x = a$ and a similarly stretched coordinate near $x = b$.

If $f(a) = f(b) = 0$ then $a_n << 1/n$ as $n \to \infty$. However, further integration by parts in (3.28) show that if the Sturm-Liouville problem is nonsingular and if $h(a)$ or $h(b) \neq 0$, then a_n behaves like $\frac{1}{n^3}$ as $n \to \infty$. In general, unless $f(x)$ satisfies an infinite number of very special conditions at $x = a$ and $x = b$, a_n decays algebraically as $n \to \infty$.

These results on algebraic decay of errors in expansions based on nonsingular second-order eigenvalue problems generalize to higher-order eigenvalue problems. For example, the expansion coefficients in a_n in $\bar{f}(x) = \sum_{n=0}^{\infty} a_n \phi_n(x)$, where $\{\phi_n(x)\}$ are the beam functions defined by

$$\phi_n'''' = \lambda_n \phi_n \,, \quad \phi_n(\pm 1) = \phi_n'(\pm 1) = 0 \,,$$

behave like $\frac{1}{n}$ if $f(\pm 1) \neq 0$ (implying a Gibbs phenomenon at the boundaries $x = \pm 1$), like $\frac{1}{n^2}$ if $f(\pm 1) = 0$ but $f'(\pm 1) \neq 0$, like $\frac{1}{n^5}$ if $f(\pm 1) = f'(\pm 1) = 0$ but $f''''(\pm 1) \neq 0$, and so on.

Singular Sturm-Liouville problems

If $p(a) = 0$ in (3.30) then it is not necessary to require that $f(a) = 0$ to achieve $a_n << \frac{1}{\lambda_n}$ as $n \to \infty$. For this reason, a Sturm-Liouville problem that is singular at $x = a$ does not lead to the Gibbs phenomenon at $x = a$. Furthermore, if the argument that led to (3.30) can be repeated on $h(x)$ given by (3.29([this is possible if p/w, p'/w, and q/w are bounded and all derivatives of f are square integrable with respect to w] then the boundary contribution to a_n from $x = a$ is smaller than $\frac{1}{\lambda_n^2}$ as $n \to \infty$. If there are no boundary contributions from $x = b$ when the operations leading to (3.30) are repeated indefinitely [which is true if $p(b) = 0$], then a_n decreases more rapidly than any power of $\frac{1}{\lambda_n}$ as $n \to \infty$.

Fourier-Bessel series

A Fourier-Bessel series of order 0 is obtained by choosing the expansion functions to be the eigenfunctions of the singular Sturm-Liouville problem

$$\frac{d}{dx}x\frac{d\phi_n}{dx} + \lambda_n x\phi_n = 0 \qquad (0 < x < 1) \qquad (3.34)$$

$$\phi_n(1) = 0, \ \phi_n(0) \text{ finite}$$

Therefore, $p(x) = w(x) = x$ in (3.25) so the problem is singular at $x = 0$, but nonsingular at $x = 1$. The eigenfunctions are

$$\phi_n(x) = J_0(j_{on}x)$$

where J_0 is the Bessel function of order 0 and j_{on} is its nth zero, $J_0(j_{on}) = 0$. The eigenvalues $\lambda_n = j_{on}^2$ satisfy

$$j_{on} \sim (n - \tfrac{1}{4})\pi \qquad (n\to\infty) .$$

The Fourier-Bessel expansion of a function $f(x)$ is given by

$$g(x) = \sum_{n=1}^{\infty} a_n J_0(j_{on}x) \tag{3.35a}$$

where (3.27) implies that

$$a_n = \frac{2}{J_0'(j_{on})^2} \int_0^1 t f(t) J_0(j_{on}t)\,dt \tag{3.35b}$$

since

$$\int_0^1 t\, J_0(j_{on}t)^2\,dt = \tfrac{1}{2} J_0'(j_{on})^2 .$$

For example, the Fourier-Bessel expansion of $f(x) = 1$ is

$$1 = -\sum_{n=1}^{\infty} \frac{2}{j_{on}J_0'(j_{on})} J_0(j_{on}x) \tag{3.36}$$

In Figure 3.6 we plot the 10, 20, and 40 term partial sums of the series (3.36). There are three noteworthy features of these plots that we will discuss:

(i) At $x = 1$ there is apparently a Gibbs phenomenon. In fact, it is easy to show that this Gibbs phenomenon has the same structure as that for Fourier sine series:

$$-\sum_{n=1}^{N} \frac{2}{j_{on}J_0'(j_{on})} J_0(j_{on} - \frac{\pi z j_{on}}{N+\frac{1}{2}}) \sim 1 + \frac{2}{\pi} \mathrm{Si}(z) \qquad (N\to\infty)$$

Since $J_0(z) \sim (2/\pi z)^{\frac{1}{2}} \cos(z-\frac{1}{4}\pi)$ as $z \to +\infty$, the large n behavior of (3.36) can be asymptotically approximated by that of Fourier series.

(ii) For fixed x satisfying $0 < x < 1$,

$$1 + \sum_{n=0}^{N} \frac{2}{j_{on}J_0'(j_{on})} J_0(j_{on}x) = 0(\tfrac{1}{N}) \qquad (N\to\infty)$$

In fact, the $\underline{n}$th term of the series has magnitude of order $1/n$ and oscillates in sign roughly every $\min(\frac{1}{x}, \frac{1}{1-x})$ terms. The error in such an oscillating series is of order $1/N$ after N terms.

(iii) At $x = 0$, the series converges (so there is no Gibbs phenomenon there) but the convergence is very slow and oscillating. In fact, the error after N terms is of order $(-1)^{N+1}/\sqrt{N}$. This follows because

$$1 + \sum_{n=0}^{N} \frac{2}{j_{on}J_0'(j_{on})} \sim \sqrt{2} \sum_{n=N+1}^{\infty} \frac{(-1)^n}{\sqrt{n}} \sim \frac{(-1)^{N+1}}{\sqrt{2N}} . \qquad (N\to\infty)$$

(3.37)

This slow rate of convergence near $x = 0$ holds even though the eigenvalue problem is singular at $x = 0$. There are two reasons why Fourier-Bessel series converge slowly near $x = 0$. First, the Gibbs phenomenon at $x = 1$ affects the rate of convergence throughout $0 \le x \le 1$. In fact, this is the sole source of the behavior (3.37). When $f'(x) \neq 0$, slow convergence near $x = 0$

can also originate from the property that $p(x) = w(x) = x$ gives $p'/w = 1/x$ which is singular at $x = 0$ so $h(x)$ given by (3.29) is singular at $x = 0$ if $f'(0) \neq 0$.

Chebyshev series revisited

The Chebyshev polynomials are the eigenfunctions of the singular Sturm-Liouville problem (3.25) with $p(x) = \sqrt{1-x^2}$, $w(x) = 1/\sqrt{1-x^2}$, $q(x) = 0$, $-1 \leq x \leq 1$, and the boundary conditions $\phi_n(\pm 1)$ finite. The eigenvalue corresponding to $T_n(x)$ is $\lambda_n = n^2$. Sine $p/w = 1-x^2$ and $p'/w = -x$ are both finite for $|x| \leq 1$, it follows that the argument leading from (3.27) to (3.30) can be repeated on $h(x)$ given by (3.29) so long as $f(x)$ is sufficiently differentiable. Therefore, the Chebyshev series expansion of an infinitely differentiable function converges faster than any power of $1/n$ as $n \to \infty$, as shown above by a different method.

To illustrate the convergence properties of Chebyshev series expansions, we study the rate of convergence of the series

$$\sin M\pi(x+a) = 2 \sum_{n=0}^{\infty} \frac{1}{c_n} J_n(M\pi) \sin(M\pi a + \tfrac{1}{2}n\pi) T_n(x) \qquad |x| \leq 1 \tag{3.38}$$

Since $J_n(M\pi) \to 0$ exponentially fast with n for $n > M$, it follows that (3.38) starts converging very rapidly when more than M terms are included (see Figure 3.7), and the conventional interpretation of these results is based on the fact that $\sin M\pi(x+a)$ has M complete wavelengths lying within $|x| \leq 1$. Thus, in order for Chebyshev expansions to converge rapidly it is necessary to retain at least π polynomials per wavelength (see Orszag & Israeli, 1974 for a similar discussion of finite difference methods).

Legendre series

Legendre polynomials are the eigenfunctions of the singular Sturm-Liouville problem (3.25) with $p(x) = 1-x^2$, $q(x) = 0$, $w(x) = 1$ for $-1 \leq x \leq 1$ and the boundary conditions are $\lambda_n = n(n+1)$ and its eigenfunction is $\lambda_n(x) = P_n(x)$, the Legendre polynomial of degree n. Since $p/w = 1 = x^2$ and $p'/w = -2x$ are both finite for $|x| \leq 1$, it follows that the Legendre series expansion of infinitely differentiable functions converges faster than algebraically.

To illustrate the convergence properties of Legendre series, we study the convergence of the series

$$\sin M\pi(x+a) = \frac{1}{\sqrt{2M}} \sum_{n=0}^{\infty} (2n+1) J_{n+\frac{1}{2}}(M\pi) \sin(M\pi a + \tfrac{1}{2}n\pi) P_n(x) \tag{3.39}$$

It follows from (3.39) that Legendre polynomial expansions of smooth functions converge rapidly provided that at least π polynomials are retained per wavelength. (See Figure 3.8)

When a discontinuous function is expanded in Legendre series, the rate of convergence is no longer faster than algebraic. In the neighborhood of a discontinuity, a Gibbs phenomenon occurs whose local structure is the same as that for Fourier series with a suitable stretching of the coordinate. For example, the Legendre series expansion of the sign function $\operatorname{sgn} x$ is

$$\operatorname{sgn} x = \sum_{n=0}^{\infty} \frac{(-1)^n (4n+3)(2n)!}{2^{2n+1}(n+1)!n!} P_{2n+1}(x) \tag{3.40}$$

The partial sums of this series are plotted in Figure 3.9. Three features are noteworthy:

(i) The Gibbs phenomenon near $x = 0$ has the same structure as that for Fourier series.

(ii) The error after N terms behaves like $1/N$ for $|x|<1$, $x \neq 0$. This follows from the fact that the $(2n+1)$st Legendre coefficient in (3.40) satisfies

$$a_n = (-1)^n \frac{(4n+3)(2n)!}{2^{2n+1}(n+1)!n!} = 0(\frac{1}{\sqrt{n}}) \qquad (n\to\infty) \qquad (3.41)$$

and the estimate

$$P_n(x) = 0(\frac{1}{\sqrt{n}}) \qquad (n \to \infty)$$

for $|x| < 1$; the series (3.40) is an alternate series if x is fixed away from zero so the error after N terms is at most order $(\frac{1}{\sqrt{n}})^2$.

(iii) The series converges only like $1/\sqrt{N}$ at $x = \pm 1$. This follows from (3.41) because $P_n(\pm 1) = (\pm 1)$ for all n. Thus, an interior Gibbs phenomenon in a Legendre series expansion has a 'long-range' effect in the sense that it seriously affects the rate of convergence at the endpoints $x = \pm 1$ of the interval.

Laguerre expansions

Laguerre polynomials are the eigenfunctions of (3.25) with $p(x) = xe^{-x}$, $q(x) = 0$, $w(x) = e^{-x}$ for $0 \leq x < \infty$ with $e^{-\frac{1}{2}x}\phi_n(x)$ bounded at $x = 0$ and ∞. The n^{th} eigenvalue is $\lambda_n = n$ and

the associated eigenfunction is $\lambda_n(x) = L_n(x)$, the Laguerre polynomial of degree n. If $f(x)$ and all its derivatives are smooth and satisfy

$$f(x) = 0(e^{\alpha x}) \qquad (x \to \infty)$$

for some $\alpha < \frac{1}{2}$, it is easy to show by retracing the derivation of (3.30) from (3.27) that the Legendre expansion

$$f(x) = \sum_{n=0}^{\infty} a_n L_n(x)$$

converges faster than algebraically as the number of terms $N \to \infty$.

To illustrate the rate of convergence of Laguerre series, we consider the expansion of $\sin x$:

$$\sin x = \sum_{n=0}^{\infty} \frac{1}{2^{(n+1)/2}} \cos \tfrac{\pi}{4}(n+1)\, L_n(x) \qquad (3.42)$$

which converges for all x, $0 \le x < \infty$. Since

$$L_n(x) \sim \frac{1}{\sqrt{\pi}} e^{\frac{1}{2}x} x^{-\frac{1}{4}} n^{-\frac{1}{4}} \cos[2\sqrt{nx} - \tfrac{1}{4}\pi] \ , \qquad (3.43)$$

[see Erdelyi et al 1953, Vol. II, pg. 200] it follows that if $N >> x$, then the error after N terms at x is roughly

$$\frac{e^{\frac{1}{2}x}}{2^{N/2}(Nx)^{\frac{1}{4}}}$$

This error is small only if $N \ell n\, 2 > x$ or $N \gtrsim 1.44x$. Since the wavelength of $\sin x$ is 2π, Laguerre expansions require approximately 9.06 polynomials per wavelength to achieve high

accuracy. (This figure may be reduced to about 6.53 polynomials per wavelength by using the modified Laguerre expansion $\sum a_n L_n(x)e^{-\alpha x}$ and optimizing the choice of α.) Thus, Laguerre expansions require many more terms to resolve a function of given complexity than do other Chebyshev or Legendre expansions. The reason is that significant weight is given to $x \to +\infty$ in the Laguerre series where $\sin x$ has an essential singularity.

In Figures 3.10-12, we plot the partial sums of (3.42) with $N = 10$, 20, and 40 terms. Observe that the number of wavelengths of $\sin x$ represented accurately by (3.42) is roughly $N/9$.

Hermite expansions

Hermite polynomials satisfy (3.25) with $p = e^{-x^2}$, $q(x) = 0$, $w(x) = e^{-x^2}$ for $-\infty < x < \infty$, $\phi_n(x)e^{-\frac{1}{2}x^2}$ bounded as $|x| \to \infty$. The Hermite polynomial $H_n(x)$ of degree n is associated with the eigenvalue $\lambda_n = 2n$. If $f(x)$ and all its derivatives satisfy

$$f(x) = 0(e^{\alpha x^2}) \qquad (|x| \to \infty)$$

for some $\alpha < \frac{1}{2}$, then the Hermite expansion

$$f(x) = \sum_{n=0}^{\infty} a_n H_n(x)$$

converges faster than algebraically as the number of terms $N \to \infty$. This is proved by retracing the steps leading from (3.27) to (3.30).

To study the rate of convergence of Hermite series, we consider the expansion of $\sin x$:

$$\sin x = \sum_{n=0}^{\infty} \frac{1}{2^{2n+1}(2n+1)!} H_{2n+1}(x) \qquad (3.44)$$

Since the asymptotic behavior of $H_n(x)$ is given by [Erdelyi, et. al 1953, vol. II, pg. 201]

$$H_n(x) \sim e^{\frac{1}{2}x^2} \frac{n!}{(\frac{1}{2}n)!} \cos\ (\sqrt{2n+1}\ x - \tfrac{1}{2}n\pi)$$

as $n \to \infty$ for x fixed, it follows that this error after N terms of (3.44) goes to zero rapidly at x only if $N \gtrsim \frac{x^2}{\log x}$. This result is very bad; to resolve M wavelengths of $\sin x$ requires nearly M^2 Hermite polynomials! [By expanding in the series $\sum a_n H_n(x) e^{-\alpha x^2}$ and optimizing the choice of α, it is possible to reduce the number of required Hermite polynomials to about $\frac{5}{2}\pi \doteq 7.85$ per wavelength, but this is still quite poor.]

Because of the poor resolution properties of Laguerre and Hermite polynomials the authors doubt they will be of much practical value in applications of spectral methods.

4. Review of Convergence Theory

The fundamental problem of the numerical analysis of initial value problems is to find conditions under which $u_N(x,t)$ converges to $u(x,t)$ as $N \to \infty$ for some time interval $0 \le t \le T$ and to estimate the error $\|u - u_N\|$. The principal result is the Lax-Richtmyer equivalence theorem which states that stability is equivalent to convergence for consistent approximations to well-posed linear problems. The terms stable, convergent, and consistent relate to technical properties of the approximation scheme which are defined below.

An approximation scheme (2.5-6) is stable if

$$\|e^{L_N t}\| \le K(t) \tag{4.1}$$

for all N where $K(t)$ is a finite function of t. Here the operator norm is defined by

$$\|e^{L_N t}\| = \max_{u \in H} \frac{\|e^{L_N t} u\|}{\|u\|} .$$

An approximation scheme is convergent if

$$\|u(t) - u_N(t)\| \to 0 \quad \text{as} \quad N \to \infty$$

for all t in the interval $0 \le t \le T$ and all $u(0) \in H$ and $f(t) \in H$. Finally, an approximation scheme is consistent if

$$\begin{aligned} \|Lu - L_N u\| &\to 0 \\ \|u - P_N u\| &\to 0 \end{aligned} \tag{4.2}$$

as $N \to \infty$ for all u in a dense subspace of H.

The foregoing definitions are standard and the Lax-Richtmyer theorem relating them is very well known (Richtmyer & Morton 1967). In this monograph we are confronted with some subtleties in the application of these ideas which will require some extensions of the notions of stability and convergence. In order to motivate these extensions, we outline here the proof of the Lax-Richtmyer theorem.

To show that stability implies convergence we use (2.1) and (2.5) to obtain

$$\frac{\partial u-u_N}{\partial t} = L_N(u-u_N) + Lu - L_N u + f - f_N$$

so

$$u(t) - u_N(t) = e^{L_N t}[u(0)-u_N(0)]$$

$$+ \int_0^t e^{L_N(t-s)}[Lu(s)-L_N u(s)+f(s)-f_N(s)]\, ds \qquad (4.3)$$

Using (4.1) and (4.3), we get

$$\|u(t)-u_N(t)\| \leq K(t)\|u(0)-u_N(0)\|$$

$$+ \int_0^t K(t-s)\left[\|Lu(s)-L_N u(s)\| + \|f(s)-f_N(s)\|\right] ds \qquad (4.4)$$

Thus, if $u(t)$ belongs to the dense subspace of H satisfying (4.2) and if $f(t)$ belongs to the dense subspace of H satisfying $\|f - P_N f\| \to 0$ as $N \to \infty$, then $\|u(t) - u_N(t)\| \to 0$

as $N \to \infty$. Since all solutions $u(t)$ of (2.1) can be approximated arbitrarily well by functions satisfying (4.2), the proof that stability implies convergence is completed.

Conversely, to show that convergence implies stability, we first observe that, for any $u \in H$, $\|e^{L_N t}u\|$ is bounded for all N and each fixed t. In fact, convergence implies

$$0 \leq \left| \|e^{L_N t}u\| - \|e^{Lt}u\| \right| \leq \|e^{L_N t}u - e^{Lt}u\| \to 0, \quad (N \to \infty)$$

while well-posedness requires that $\|e^{Lt}u\|$ is finite. However, $\max_N \|e^{L_N t}u\|$ may depend on u and on t, so stability is not yet proved. To complete the proof we use the fact that H is a Hilbert space. The principle of uniform boundedness (Richtmyer & Morton 1967) implies that if $\|e^{L_N t}u\|$ is bounded as $N \to \infty$ for each t and $u \in H$ then $\|e^{L_N t}\|$ is bounded as $N \to \infty$ for each t. This proves stability.

Using the Lax-Richtmyer theorem, the study of the convergence of discrete approximations to the solutions of initial-value problems is reduced to the study of the stability of the discrete approximations, assuming the approximations are consistent. Thus, the development of conditions for the stability of families of finite-dimensional operators L_N is of primary interest in numerical analysis.

The simplest condition for stability is due to von Neumann. Let us suppose that the Hilbert space H possesses the inner product $(,)$. If each L_N is a normal operator [that is, the adjoint L_N^* defined with respect to $(,)$ commutes with

L_N so $L_N L_N^* = L_N^* L_N$] then stability is equivalent to the von Neumann condition

$$\mathrm{Re}\lambda_N < C \tag{4.7}$$

where λ_N is any of the eigenvalues of any of the operators L_N and C is a finite constant independent of N. To prove this, we note that if L_N is normal, then L_N and L_N^* as well as $\exp(L_N t)$ and $\exp(L_N^* t)$, are simultaneously diagnolizable. Therefore,

$$\|e^{L_N t}\|^2 = \max_{u \in H} \frac{(u, e^{L_N^* t} e^{L_N t} u)}{(u,u)} = \max_{\lambda_N} e^{2(\mathrm{Re}\lambda_N)t}$$

where λ_N are the eigenvalues of L_N. Thus, the von Neumann condition (4.7) is equivalent to the stability definition (4.1) with $K(t) = \exp(2Ct)$.

The von Neumann condition gives an operational technique for checking stability of normal approximations: compute the eigenvalues of L_N and check that the real parts of the eigenvalues are bounded from above.

Example 4.1: Symmetric hyperbolic system with periodic boundary conditions

Let us apply the theory just discussed to the stability of difference approximations to the m-component symmetric hyperbolic system

$$\frac{\partial \vec{u}(x,t)}{\partial t} = A \frac{\partial \vec{u}(x,t)}{\partial x} \tag{4.8}$$

with periodic boundary conditions $\vec{u}(0,t) = \vec{u}(1,t)$.

Here $\vec{u}$ is an m-component eigenvector and A is a symmetric $m \times m$ matrix.

If we discretize in space using second-order centered differences, we obtain

$$\frac{\partial \vec{u}_j}{\partial t} = A \frac{\vec{u}_{j+1} - \vec{u}_{j-1}}{2\Delta x} \qquad (j = 1,2,\ldots,N) \tag{4.9}$$

$$\vec{u}_0(t) = \vec{u}_N(t) \ , \quad \vec{u}_1(t) = \vec{u}_{N+1}(t)$$

where $\vec{u}_k(t) = u(k/N,t)$ and $\Delta x = 1/N$. The system (4.9) is equivalent to the system of mN equations

$$\frac{\partial \hat{u}}{\partial t} = B \hat{u} \tag{4.10a}$$

where $\hat{u}^T = (\vec{u}_1,\vec{u}_2,\ldots,\vec{u}_N)$ and B is the $mN \times mN$ matrix given as the Kronecker product

$$B = A \otimes D \tag{4.10b}$$

where A is the $m \times m$ matrix in (4.8) and D is the $N \times N$ matrix

$$D = \frac{1}{2\Delta x} \begin{pmatrix} 0 & 1 & 0 & 0 & \ldots & 0 & -1 \\ -1 & 0 & 1 & 0 & \ldots & 0 & 0 \\ 0 & -1 & 0 & 1 & \ldots & 0 & 0 \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & & \cdot & \cdot \\ 1 & 0 & 0 & 0 & & -1 & 0 \end{pmatrix} .$$

D is anti-symmetric (and, hence, normal) so it has eigenvalues 0 and pure imaginary. In fact, the eigenvalues of D are $i \sin(2\pi k\Delta x)/\Delta x$ for $k = 0,1,\ldots,N-1$. Thus, the norm of $\exp(Bt)$ satisfies

$$\|\exp(Bt)\| = \max_{0\le k<N} \|\exp(iA \sin(2\pi k\Delta x)t/\Delta x\| = 1 ,$$

where we use the fact that A is symmetric so it has real eigenvalues.

If the approximate evolution operators L_N are not normal, conditions guaranteeing stability are much harder to obtain. One important case in which stability conditions can be obtained is for the problem studied in Example 4.1 with A no longer symmetric. More generally, suppose the approximation L_N has the form $L_N = A \otimes D_N$ where A is a fixed $m \times m$ matrix (possibly not normal) and D_N is an N-dimensional normal matrix. It is easy to show that

$$\|\exp(L_N t)\| = \max_{\lambda_N} \|\exp(\lambda_N A t)\| \qquad (4.11)$$

where λ_N is any of the eigenvalues of D_N.

To investigate the stability of $\exp(L_N t)$ we generalize (4.11) further and seek conditions for the stability of a family of $m \times m$ matrices $A(\omega)$, where ω is an arbitrary parameter. That is, we seek conditions such that

$$\max_{\omega} \|\exp[A(\omega)t]\| \le K(t) ,$$

where $K(t)$ is a finite function of t. Once these general conditions are found, they can be specialized to give stability conditions for families of the form $\exp(L_N t)$ where $L_N = A \otimes D_N$ with D_N normal by choosing $A(\omega) = A\omega$ where ω is any of the eigenvalues of any of the matrices D_N

The basic result on the stability of families of $m \times m$ matrices is the Kreiss matrix theorem (Kreiss 1962):

For any family $A(\omega)$ of $m \times m$ matrices, each of the following statements implies the next:

(i) There exist symmetric matrices $H(\omega)$ satisfying $H(\omega)A(\omega) + A^*(\omega)H(\omega) \leq 0$ and $I \leq H(\omega)$, $\|H(\omega)\| \leq C$ for some constant C.

(ii) $\|\exp[A(\omega)t]\| \leq C$ for all $t \geq 0$.

(iii) $(\text{Re } \lambda)\|(\lambda I - A(\omega)^{-1}\| \leq C'$ for some constant C' and all λ satisfying $\text{Re } \lambda > 0$.

(iv) There exist matrices $H(\omega)$ satisfying (i) with $\|H(\omega)\| \leq K(m)C'$ where C' is the constant appearing in (iii) and $K(m)$ depends only on m and not only the family $A(\omega)$.

Observe that for a family of matrices $A(\omega)$ to satisfy the conditions of this theorem it is necessary that all the eigenvalues of all the matrices have non-positive real parts. Otherwise there would be some ω and some eigenvector $\vec{u}$ satisfying $\|\exp[A(\omega)t]\vec{u}\| \to \infty$ as $t \to \infty$ violating (ii).

The most important relation implied by this theorem is the implication that (iii) implies (ii) with $C \leq K(m)C'$ That is, for any $m \times m$ matrix A all of whose eigenvalues have nonpositive real parts

$$\|\exp(At)\| \leq K'(m) \max_{\text{Re } \lambda > 0} (\text{Re } \lambda)\|(\lambda I - A)^{-1}\| \qquad (4.12)$$

where $K'(m)$ is a finite function of m.

An elementary proof of (4.12) has recently been given by Lapfev (1975) and improved by C. McCarthy (private communication to G. Strang, 1975). Lapfev observes that if $\nu > 0$, then

$$e^{At} = \frac{1}{2\pi i}\int_{\nu-i\infty}^{\nu+i\infty} e^{\lambda t}(\lambda I - A)^{-1}\, d\lambda = \frac{e^{\nu t}}{2\pi}\int_{-\infty}^{\infty} e^{i\mu t}(\nu + i\mu - A)^{-1}\, d\mu , \quad (4.13)$$

as may be proved by shifting contours in the complex plane. Since each entry of $(\nu+i\mu-A)^{-1}$ is a rational function in μ of degree at most m, the derivatives of the real and imaginary parts of each entry can change sign at most $4m$ times when μ increases from $-\infty$ to ∞. On any μ-interval, say $a \le \mu \le b$, where the real and imaginary parts of an entry in $(\nu+i\mu-A)^{-1}$ are monotonic, the second mean-value theorem implies

$$\int_a^b \cos \mu t\, f(\mu)\, d\mu = f(a)\left[\frac{\sin(ct) - \sin(at)}{t}\right] + f(b)\left[\frac{\sin(bt) - \sin(ct)}{t}\right]$$

$$\le \frac{4}{t}\max_{\mu}|f(\mu)| ,$$

for some c satisfying $a < c < b$ where $f(\mu) = \mathrm{Re}(\nu+i\mu-A)^{-1}_{ij}$. Thus, for all i,j

$$\left|\int_{-\infty}^{\infty} e^{i\mu t}(\nu+i\mu-A)^{-1}_{ij}\, d\mu\right| \le \frac{64\,m}{t}\max_{\mu}\left|(\nu+i\mu-A)^{-1}_{ij}\right| . \quad (4.14)$$

If it is true that the matrix norm has the property that $|B_{ij}| \le C_{ij}$ for all i,j implies $\|B\| \le \|C\|$, then (4.14) implies

$$\left\|\int_{-\infty}^{\infty} e^{i\mu t}(\nu+i\mu-A)^{-1}\, d\mu\right\| \le \frac{64\,m}{t}\max_{\mu}\left\|(\nu+i\mu-A)^{-1}\right\| \quad (4.15)$$

Choosing $\nu = 1/t$ in (4.13-15) gives (4.12) with $K'(m) = 64\,m$.

There are three important matrix norms in which $|B_{ij}| \le C_{ij}$ for all i,j implies $\|B\| \le \|C\|$, namely the matrix norms induced by the L_1, L_2, and L_∞ vector norms. This is shown using the relations

$$\|B\|_1 = \max_j \sum_{i=1}^{m} |B_{ij}|$$

$$\|B\|_2 = \sup_{\substack{\|x\|_2=1 \\ \|y\|_2=1}} \sum_{i=1}^{m} \sum_{j=1}^{m} B_{ij} x_i y_i$$

$$\|B\|_\infty = \max_i \sum_{j=1}^{m} |B_{ij}|$$

which hold for all matrices B. In other norms $|B_{ij}| \le C_{ij}$ may not imply $\|B\| \le \|C\|$ but the equivalence of all matrix norms implies $\|B\| \le F(m)\|C\|$ for some finite function of the dimension m. Thus, (4.12) is obtained with $K'(m) = 64\, m\, F(m)$.

The functions $K(m)$ appearing in statement (iv) of the Kreiss theorem and $K'(m)$ appearing in (4.12) need not be equal. It follows from the Kreiss theorem that $K'(m) \le K(m)$. Kreiss showed only that $K(m) = 0(m^m)$ as $m \to \infty$; this is much too conservative. Miller & Strang (1965) showed that $K(m) = 0(C^m)$ as $m \to \infty$ for some constant $C > 1$.

In the case of a normal family of matrices $A(\omega)$ the conditions of the Kreiss matrix theorem are trivially satisfied: if the eigenvalues of $A(\omega)$ have negative real parts then $\|\exp[A(\omega)t]\| \le 1$ for all $t \ge 0$ and ω.

Unfortunately, the class of semi-discrete approximations investigated in this monograph does not easily fit within the class of problems to which the above stability conditions can be applied. In contrast to the classical problems of the numerical analysis of difference methods for initial-value problems, spectral approximations L_N are frequently not normal nor even approximately normal. [There is an important extension of stability analysis to non-normal approximations obtained by finite-difference approximation to mixed initial-boundary value problems. The non-normality of these problems is frequently induced by the boundary conditions and constitutes a small perturbation of a normal approximation. In this case, extensions of von Neumann stability analysis, like that introduced by Godunov and Ryabenkii (see Richtmyer & Morton 1967) apply.]

5. Algebraic Stability

In this section, we develop a theory of stability and convergence which generalizes the classical theory discussed in Sec. 4. As will be shown by examples in Sects. 6-9, this generalized stability theory is well suited to study the convergence of spectral methods.

A spectral approximation

$$\frac{\partial u_N}{\partial t} = L_N u_N + f_N \tag{5.1}$$

to the initial-value problem $u_t = Lu + f$ is called <u>algebraically stable</u> as $N \to \infty$ if

$$\| e^{L_N t} \| \leq N^r N^{st} K(t) \tag{5.2}$$

for all sufficiently large N, where r, s, and $K(t)$ are finite for $0 \leq t \leq T$.

It may at first seem that the Lax-Richtmyer theorem shows that algebraically stable approximations cannot be convergent unless (5.2) holds with $r \leq 0$, $s \leq 0$. In fact, if we demand that the approximations converge for all $u(0)$ and $f(t)$ in the Hilbert space $\mathcal{H}$, this conclusion <u>is</u> correct. However, it is possible for approximations that satisfy (5.2) with $r > 0$ or $s > 0$ to converge on a dense subset of the Hilbert space in which the only functions for which convergence is not obtained are highly pathological. In fact, if $p = r + sT > 0$ but p is smaller than the order of the

spatial truncation error of a particular solution $u(x,t)$, i.e.

$$N^p \, \| Lu(t) - L_N u(t) \| \to 0 \qquad (N \to \infty) \qquad (5.3a)$$

$$N^p \, \| u(0) - u_N(0) \| \to 0 \qquad (N \to \infty) \qquad (5.3b)$$

$$N^p \, \| f(t) - f_N(t) \| \to 0 \qquad (N \to \infty) \qquad (5.3c)$$

for all $0 \le t \le T$, then (4.4) and (5.2) imply that

$$\| u(t) - u_N(t) \| \to 0 \qquad (N \to \infty)$$

for $0 \le t \le T$. Thus, algebraic stability implies convergence in that subspace of $\mathcal{H}$ satisfying the conditions (5.3). If this latter subspace is large enough, an algebraically stable method can still be very useful although it cannot yield convergent results for all initial conditions $u(0)$ and forces $f(t)$. Since spectral methods are normally infinite-order accurate, algebraic stability implies convergence for such spectral methods.

In the examples of algebraic stability given in Sects. 7-9, we find $r \le \frac{1}{4}$, $s \le 0$, and $K(t) \le M$. In this case, algebraic stability implies convergence so long as (5.3) holds with $p \le \frac{1}{4}$. Thus, the approximation need not be infinite-order accurate to achieve convergence. However, we develop the general theory of algebraic stability here in the expectation that it will find application to spectral methods for high-order equations in which p may be large.

Our definition of algebraic stability is very similar to the notion of s-stability introduced by Strang (1960). However, our motivation is slightly different. Strang introduced s-stability to study the convergence of time-discretized initial-value problems in which the norm of the evolution operator grows as a power of the time step. We shall return to this concept of s-stability in Sec. 10.

Let us give an illustration of the need for a theory of algebraic stability. In Sec. 8, we will discuss Chebyshev polynomial spectral methods to solve the one-dimensional wave equation $u_t + u_x = f(x,t)$ with boundary conditions $u(-1,t) = 0$. Unfortunately this problem is <u>not</u> well posed in the Chebyshev norm

$$\|u\|^2 = \int_{-1}^{1} \frac{u^2(x)}{\sqrt{1-x^2}}\, dx ;$$

in fact, if

$$u(x,0) = \begin{cases} 1 - \dfrac{|x|}{\epsilon} & \text{if } |x| < \epsilon \\ 0 & \text{if } |x| \geq \epsilon \end{cases}$$

then the solution of $u_t + u_x = 0$, $u(-1,t) = 0$ at $t = 1$ is given by

$$u(x,1) = \begin{cases} 1 - \dfrac{1}{\epsilon} + \dfrac{x}{\epsilon} & 1-\epsilon < x \leq 1 \\ 0 & x \leq 1-\epsilon \end{cases}$$

Therefore, as $\varepsilon \to 0+$,

$$\|u(x,0)\|^2 \sim \varepsilon \qquad (\varepsilon \to 0+)$$

$$\|u(x,1)\|^2 \sim \frac{2}{3}\sqrt{2\varepsilon} \qquad (\varepsilon \to 0+) ,$$

so that if $L = -\frac{\partial}{\partial x}$,

$$\|e^L\| \geq \frac{\|u(x,1)\|}{\|u(x,0)\|} \sim \left(\frac{8}{9}\right)^{\frac{1}{4}} \varepsilon^{-\frac{1}{4}} \qquad (\varepsilon \to 0+) \qquad (5.4)$$

In fact, $\|e^{Lt}\| = \infty$ for $0 < t < 2$, $\|e^{Lt}\| = 0$ for $t > 2$, so the one-dimensional wave equation is not well posed in the Chebyshev norm.

Since the finite-dimensional approximations L_N to L given by Galerkin, tau, and collocation approximation (see Sec. 2) should converge as $N \to \infty$, it follows that we may expect

$$\|\exp(L_N t)\| \to \infty$$

as $N \to \infty$ in the Chebyshev norm. To estimate the rate of divergence of $\|\exp(L_N t)\|$ as $N \to \infty$ we argue that Chebyshev polynomials of degree at most N can resolve distances of at most order $1/N$ <u>interior</u> to $(-1,1)$ so we may reasonably guess on the basis of (5.4) with $\varepsilon = 1/N$ that

$$\|\exp(L_N t)\| = O\left(N^{\frac{1}{4}}\right) \qquad (N \to \infty) . \qquad (5.5)$$

This result is justified by the numerical results presented in Table 8.3.

Thus, we expect that Chebyshev-spectral approximations to the one-dimensional wave equation are not stable but are algebraically stable with $r = 1/4$ and $s = 0$ in (5.2).

Notice that algebraic stability in one norm implies algebraic stability in all algebraically equivalent norms. Thus, algebraic stability is equivalent in all of the L_p norms $1 \leq p \leq \infty$ because these norms are algebraically equivalent in N-dimensional vector spaces (i.e., they differ from each other only by a fixed power of N). To show this, we recall that the L_p norm of a vector $\vec{a} = (a_1,\ldots,a_N)$ is defined by

$$\|a\|_p = \left(\sum_{i=1}^{N} |a_i|^p\right)^{1/p} .$$

If $q = p\alpha$ with $0 < \alpha < 1$, then

$$\|a\|_q^q = \left(\sum_{i=1}^{N} |a_i|^{p\alpha}\right) \leq \left(\sum_{i=1}^{N} |a_i|^p\right)^{\alpha}\left(\sum_{i=1}^{N} 1\right)^{1-\alpha} = \|a\|_p^q \; N^{1-q/p}$$

by Holder's inequality. Therefore, for all $p > 1$,

$$N^{\frac{1}{p}-1} \|a\|_1 < \|a\|_p$$

Also, if $p > 1$, then

$$\|a\|_p^p = \sum_{i=1}^{N} |a_i|^p \leq \left(\sum_{i=1}^{N} |a_i| \right)^p = \|a\|_1^p ,$$

so that

$$N^{\frac{1}{p}-1} \|a\|_1 \leq \|a\|_p \leq \|a\|_1 . \qquad (5.6)$$

The verification of algebraic stability for spectral methods leads to a general problem in matrix theory. Suppose that $A_N (N=1,2,\ldots,)$ is a one parameter family of matrices. We will find conditions on the members of the family such that $\exp(A_N t)$ is algebraically stable. We will use only the L_2 norm since the others are equivalent to it.

Conditions for Algebraic Stability

Let $\{A_N\}$ be a family of $N \times N$ matrices where $\|A_N\| = O(N^a)$ $(N \to \infty)$ for some finite a. A necessary and sufficient condition for algebraic stability

$$\|e^{A_N t}\| = O\left(N^{r_N} s^t\right) \qquad (N \to \infty)$$

is that there exist a family $\{H_N\}$ of Hermitian positive-definite matrices such that

$$\|H_N^{-1}\| \, \|H_N\| = O(N^b) \qquad (N \to \infty) \tag{5.7a}$$

$$H_N A_N + A_N^* H_N \leq c(N) H_N \tag{5.7b}$$

$$c(N) < d \log N \tag{5.7c}$$

for all sufficiently large N where b and d are finite numbers independent of N.

To prove sufficiency we use the Lie formula

$$e^{(C+D)t} = \lim_{n\to\infty} \left(e^{Ct/n} \, e^{Dt/n} \right)^n \tag{5.8}$$

which is valid for arbitrary matrices C and D. This formula is proved at the end of this section. If we define

$$C = \frac{1}{2} \left[H_N^{\frac{1}{2}} A_N H_N^{-\frac{1}{2}} + H_N^{-\frac{1}{2}} A_N^* H_N^{\frac{1}{2}} \right]$$

$$D = \frac{1}{2} \left[H_N^{\frac{1}{2}} A_N H_N^{-\frac{1}{2}} - H_N^{-\frac{1}{2}} A_N^* H_N^{\frac{1}{2}} \right] \tag{5.9}$$

and note that

$$\exp [A_N t] = H_N^{-\frac{1}{2}} \exp \left[H_N^{\frac{1}{2}} A_N H_N^{-\frac{1}{2}} t \right] H_N^{\frac{1}{2}} ,$$

it follows from the Lie formula that

$$e^{A_N t} = \lim_{n\to\infty} H_N^{-\frac{1}{2}} \left(e^{Ct/n}\, e^{Dt/n}\right)^n H_N^{\frac{1}{2}} \qquad (5.10)$$

However, it follows from (5.7b) that, since C is a symmetric matrix,

$$\|e^{Ct/n}\| \leq e^{ct/n} .$$

Also, D is an antisymmetric matrix so that

$$\|e^{Dt/n}\| = 1$$

Therefore, (5.10) gives

$$\|e^{A_N t}\| \leq e^{ct}\, \|H_N^{-\frac{1}{2}}\|\, \|H_N^{\frac{1}{2}}\| \leq e^{ct} N^{b/2}$$

proving algebraic stability.

In order to prove that the conditions (5.7) are also necessary for algebraic stability we define

$$B_N = A_N - (r+1)\log(N) I .$$

Therefore,

$$\|e^{B_N t}\| = 0\left(\frac{N^s}{N^t}\right) \qquad (N \to \infty) .$$

By Liapounov's theorem (Barnett & Storey 1974) there exists a Hermitian positive-definite matrix H_N such that

$$H_N B_N + B_N^* H_N = -I , \tag{5.11}$$

Thus,

$$H_N A_N + A_N^* H_N = -I + 2(r+1) \log N \, H_N \leq c(N) H_N$$

where $c(N) = 2(r+1) \log N$. In order to complete the proof of (5.7) we need to estimate the norms of H_N and H_N^{-1} . It can be easily verified that an explicit formula for H_N is

$$H_N = \int_0^\infty e^{B_N t} e^{B_N^* t} \, dt .$$

Therefore,

$$\|H_N\| \leq \int_0^\infty \|e^{B_N t}\| \, \|e^{B_N^* t}\| \, dt \leq N^{2s} \int_0^\infty N^{-2t} \, dt \leq N^{2s}$$

if $2 \ell n N > 1$, i.e., $N \geq 2$. Also from (5.11) we obtain

$$B_N H_N^{-1} + H_N^{-1} B_N^* = - (H_N^{-1})^2$$

so that

$$\|H_N^{-1}\|^2 \leq 2\|B_N\| \|H_N^{-1}\|$$

or

$$\|H_N^{-1}\| \leq 2\|B_N\| = O(N^a) \qquad (N \to \infty) \qquad (5.12)$$

This completes the proof of the necessity of (5.7).

The above result gives a method for checking numerically the algebraic stability of a family $\{A_N\}$ of matrices satisfying $\|A_N\| = O(N^a)$ as $N \to \infty$:

(i) We check that the real parts of the eigenvalues of A_N are bounded from above by $s \log N$; otherwise, the family of matrices A_N are algebraically unstable.

(ii) We introduce $B_N = A_N - (s+1)\log(N) I$ and compute the Liapounov matrix H_N such that $H_N B_N + B_N^* H_N = -I$. There are several numerically efficient techniques to compute H_N (Bartels & Stewart 1972).

(iii) To verify algebraic stability the condition number of H_N must be bounded by N^b for some finite b as $N \to \infty$. Noting (5.12), it is only necessary to verify that the eigenvalues of H_N are bounded from above by some finite power of N as $N \to \infty$.

This procedure is applied in Sec. 8 to verify algebraic stability of some problems. Since (5.7) gives a necessary and sufficient condition for algebraic stability, if these conditions do not hold the family of matrices A_N is algebraically unstable.

Finally, we prove the Lie formula (5.8) for finite dimensional matrices. First, we write

$$e^{C+D} - \left(e^{\frac{C}{n}}e^{\frac{D}{n}}\right)^n = e^{\left(\frac{C+D}{n}\right)^n} - e^{\frac{C}{n}}e^{\frac{D}{n}}$$

$$= \sum_{k=0}^{n-1} e^{\left(\frac{C+D}{n}\right)^k}\left(e^{\frac{C+D}{n}} - e^{\frac{C}{n}}e^{\frac{D}{n}}\right)\left(e^{\frac{C}{n}}e^{\frac{D}{n}}\right)^{n-1-k} .$$

Therefore,

$$\left\| e^{C+D} - \left(e^{\frac{C}{n}}e^{\frac{D}{n}}\right)^n \right\| \leq \sum_{k=0}^{n-1} e^{\|C+D\|\frac{k}{n}} \left\| e^{\frac{C+D}{n}} - e^{\frac{C}{n}}e^{\frac{D}{n}} \right\|$$

$$\times \left(e^{\|C\| + \|D\|}\right)^{\frac{n-1-k}{n}}$$

$$\leq \left\| e^{\frac{C+D}{n}} - e^{\frac{C}{n}}e^{\frac{D}{n}} \right\| n\, e^{\theta\frac{n-1}{n}}$$

where

$$\theta = \|C\| + \|D\| .$$

On the other hand,

$$\|e^{\frac{C+D}{n}} - e^{\frac{C}{n}}e^{\frac{D}{n}}\| \leq \frac{\|CD-DC\|}{2n^2} + 0\left(\frac{1}{n^3}\right) \qquad (n \to \infty)$$

so that

$$\|e^{C+D} - \left(e^{\frac{C}{n}}e^{\frac{D}{n}}\right)^n\| \leq \frac{K}{n} \qquad (n \to \infty)$$

for any $K > \frac{1}{2}\|CD-DC\|$, proving (5.8).

Eq. (5.8) is also true for certain infinite dimensional matrices (operators). This deep result, called the Trotter product formula, is very useful in the modern theory of partial differential equations.

BIBLIOGRAPHY

Barnett, S. and Storey, C. (1974) MATRIX METHODS IN STABILITY THEORY, Barnes & Noble, New York.

Bartels, R. and Stewart, G. (1972), "Solution of the Matrix Equation AX + XB = C," Comm. ACM, Vol. 15, pp. 820-826.

Collatz, L. (1960) THE NUMERICAL TREATMENT OF DIFFERENTIAL EQUATIONS, 3rd ED., Springer-Verlag, Berlin.

Courant, R. and Hilbert, D. (1953) METHODS OF MATHEMATICAL PHYSICS, Part I, Interscience, New York.

Erdelyi, A. (1953) HIGHER TRANSCENDENTAL FUNCTIONS, VOL. II, McGraw-Hill, New York.

Lanczos, C. (1956) APPLIED ANALYSIS, Prentice-Hall, Englewood Cliffs, New Jersey.

Lapfer, A. (1975), Soviet Math. Dokl., Vol. 16, pp. 65-69.

Miller, J. J. H., and Strang, W. G. (1965), "Matrix Theorems for Partial Differential and Difference Equations," Stanford University Tech. Report CS28, Stanford, California.

Richtmyer, R. D. and Morton, K. W. (1967) DIFFERENCE METHODS FOR INITIAL VALUE PROBLEMS, 2nd Ed., Interscience Tracts in Pure and Applied Mathematics, No. 4, Interscience, New York.

Orszag, S. A. (1971a): "Numerical Simulation of Incompressible Flows Within Simple Boundaries: Galerkin (spectral) Representations," STUDIES IN APPLIED MATHEMATICS, Vol. 50, pp. 293-327.

Orszag, S. A. and Israeli, M. (1974): "Numerical Simulation of Viscous Incompressible Flows," Annual Review Fluid Mechanics, Vol. 5.

Strang, W. G. (1960): "Difference Methods for Mixed Boundary-Value Problems," Duke Mathematical Journal, Vol. 27, p. 221.

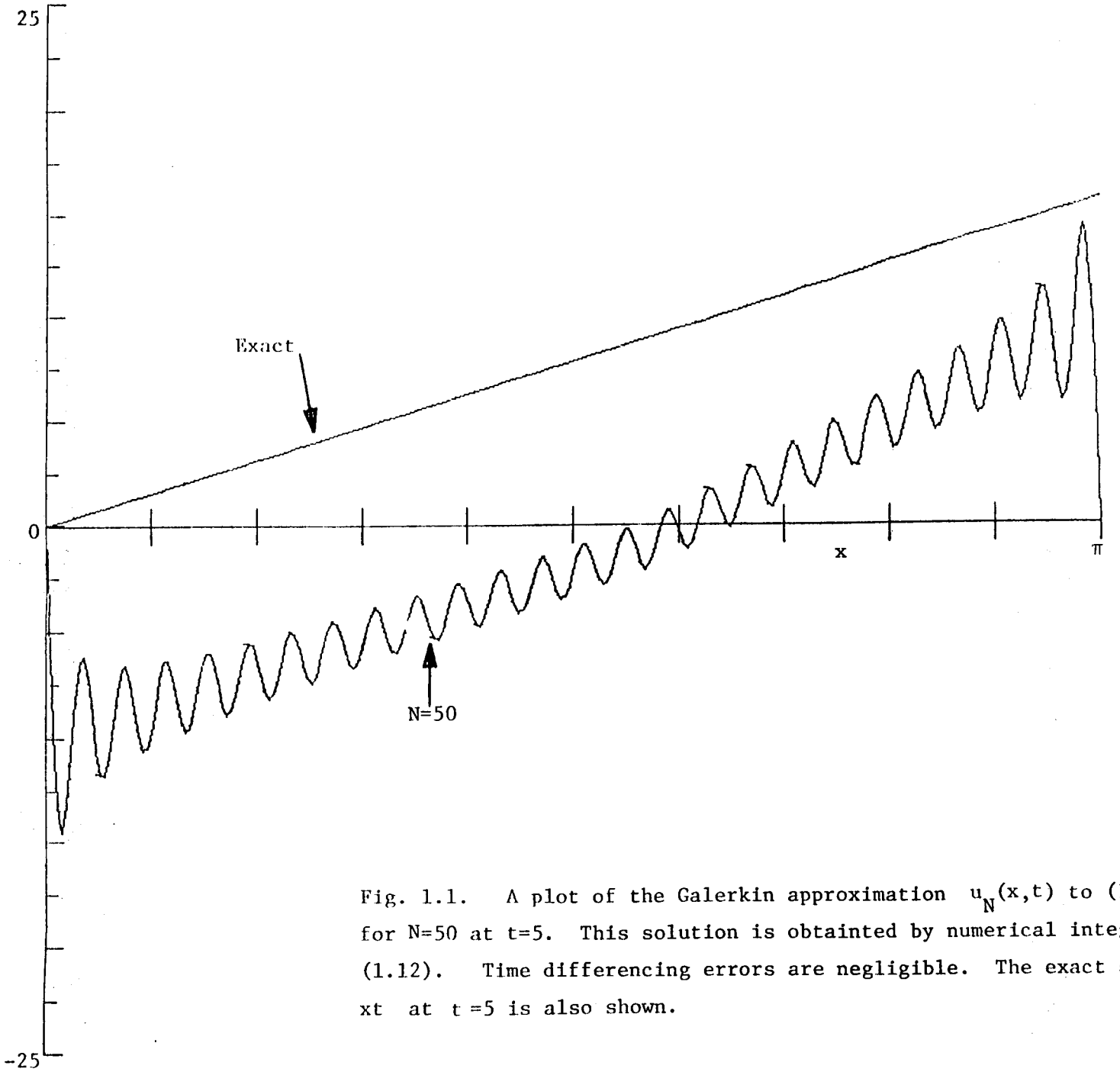

Fig. 1.1. A plot of the Galerkin approximation $u_N(x,t)$ to (1.8) for N=50 at t=5. This solution is obtainted by numerical integration of (1.12). Time differencing errors are negligible. The exact solution xt at t =5 is also shown.

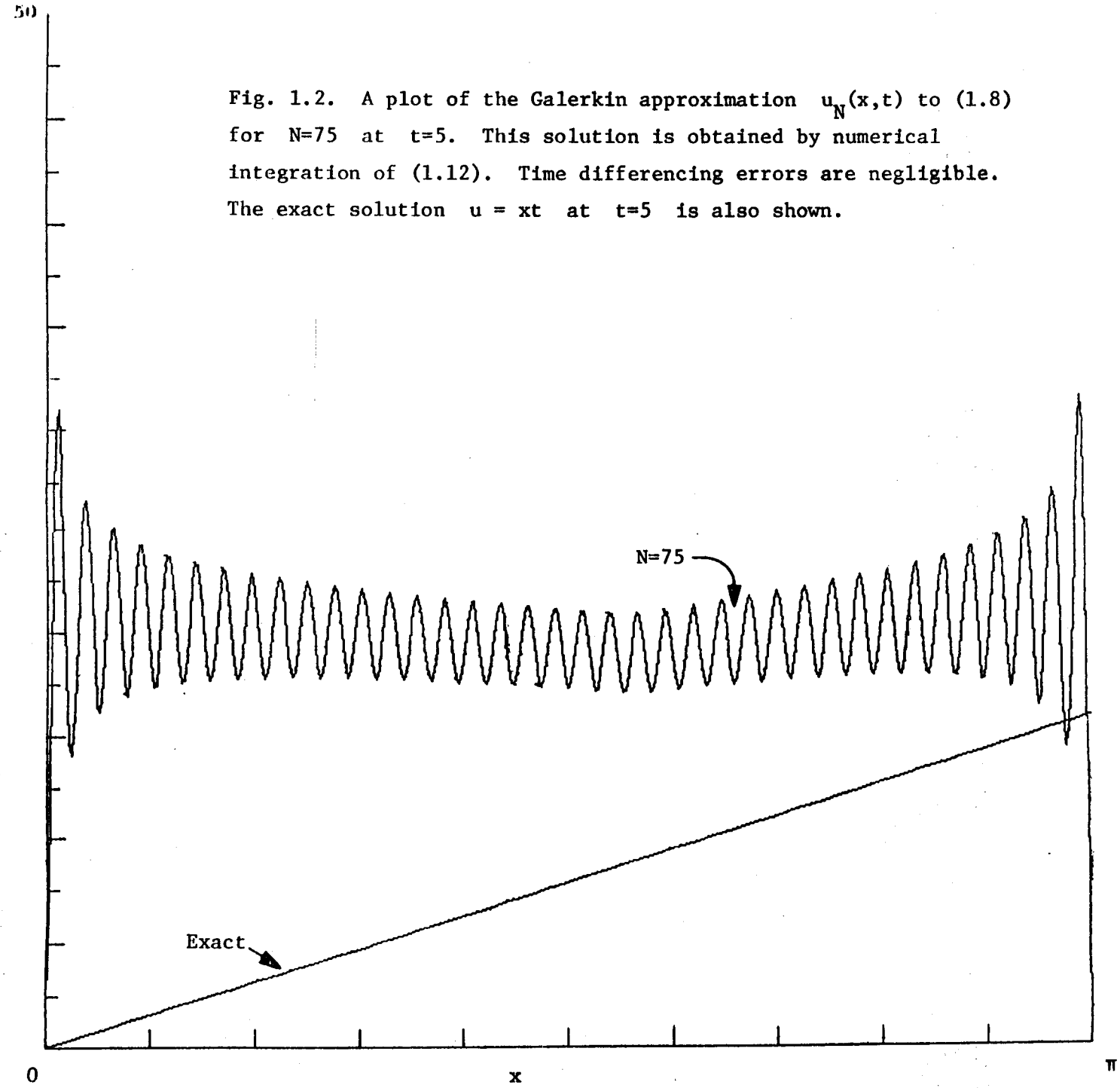

Fig. 1.2. A plot of the Galerkin approximation $u_N(x,t)$ to (1.8) for N=75 at t=5. This solution is obtained by numerical integration of (1.12). Time differencing errors are negligible. The exact solution u = xt at t=5 is also shown.

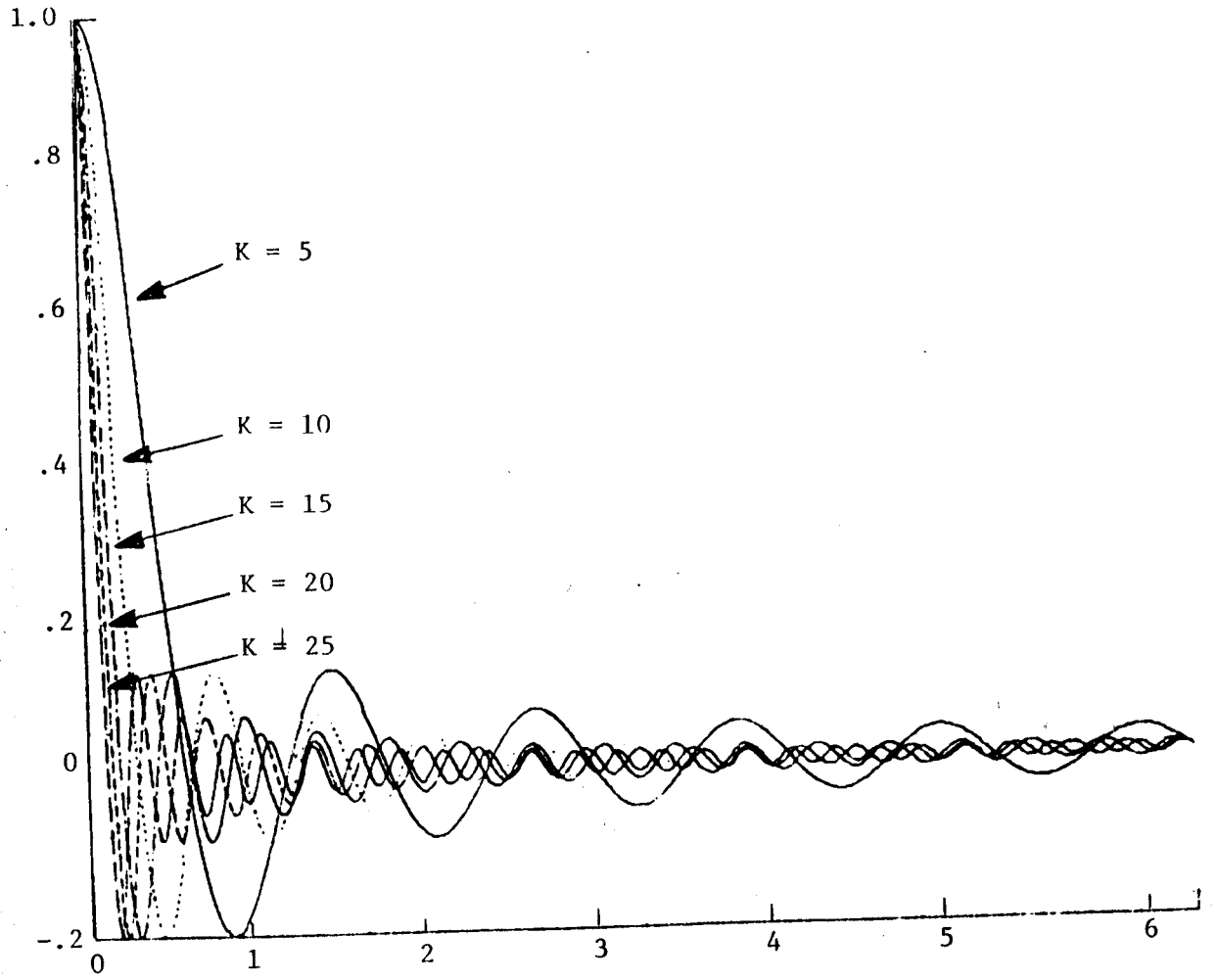

Fig. 3.1. A plot of $\sin(K+\frac{1}{2})t/[(K+\frac{1}{2})t]$ for K = 5, 10, 15, 20, 25.

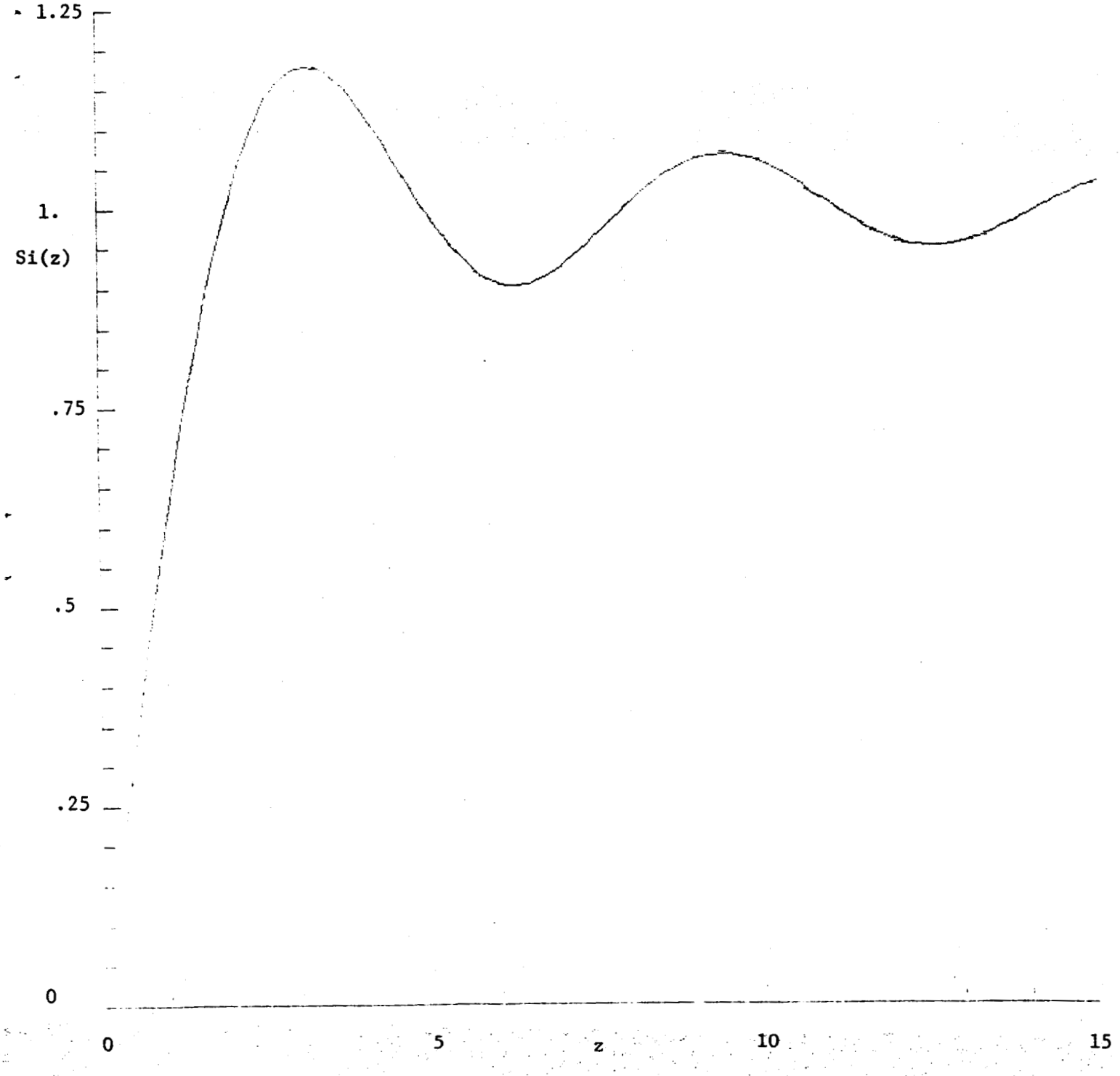

Fig. 3.2. A plot of the sine integral Si(z) defined in (3.14b) for $0 \leq z \leq 15$.

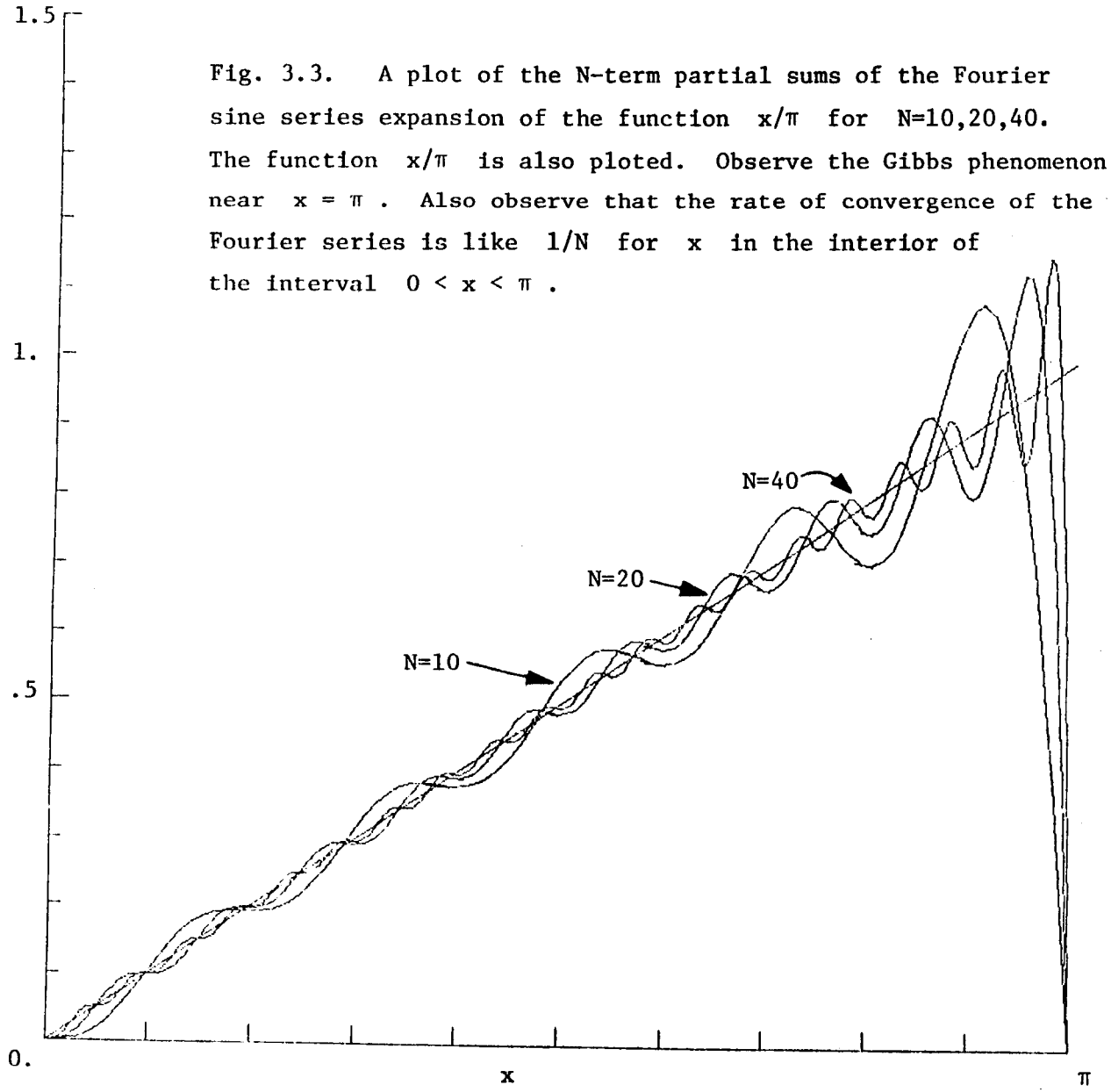

Fig. 3.3. A plot of the N-term partial sums of the Fourier sine series expansion of the function x/π for N=10,20,40. The function x/π is also ploted. Observe the Gibbs phenomenon near $x = \pi$. Also observe that the rate of convergence of the Fourier series is like $1/N$ for x in the interior of the interval $0 < x < \pi$.

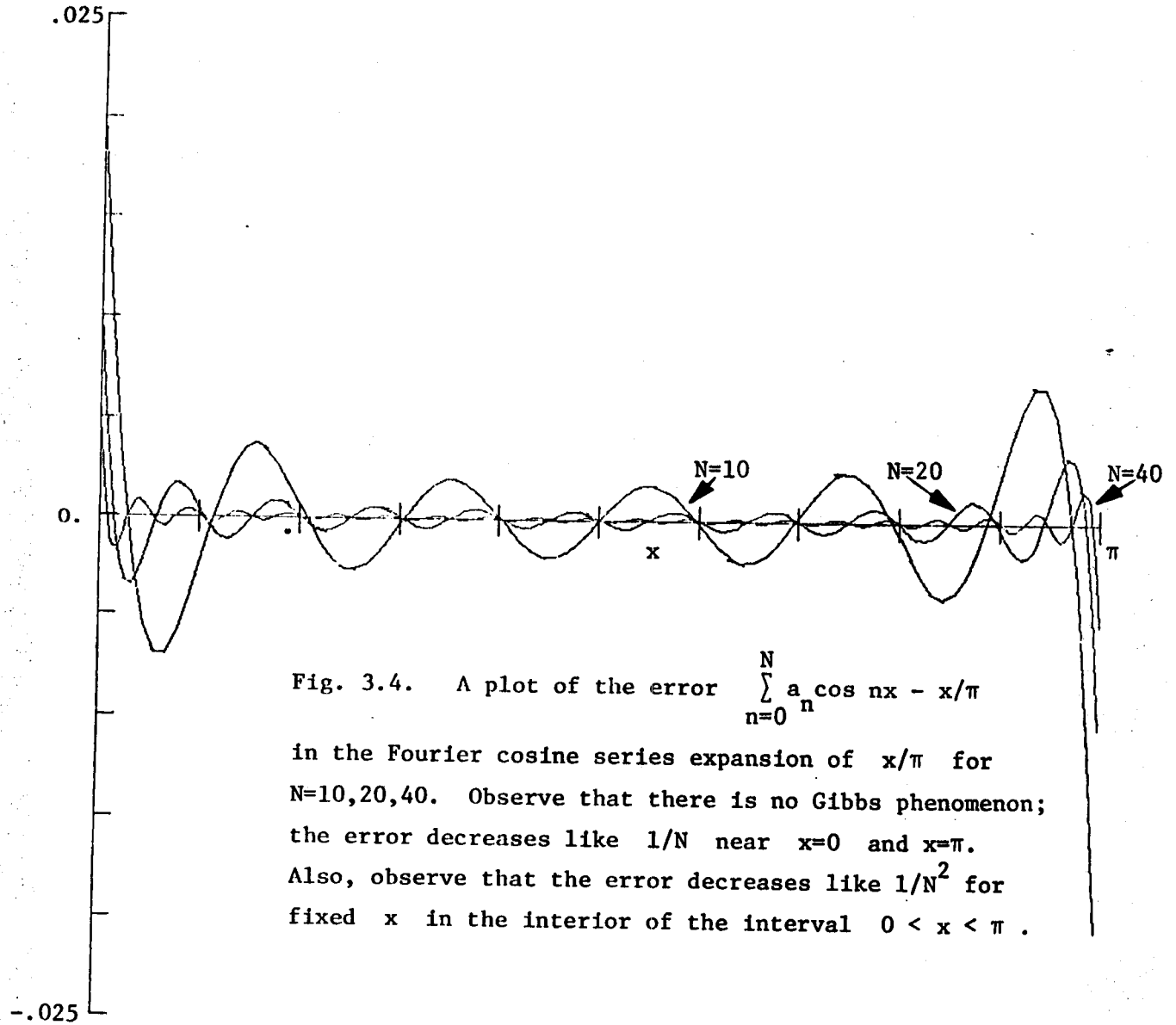

Fig. 3.4. A plot of the error $\sum_{n=0}^{N} a_n \cos nx - x/\pi$ in the Fourier cosine series expansion of x/π for N=10,20,40. Observe that there is no Gibbs phenomenon; the error decreases like $1/N$ near $x=0$ and $x=\pi$. Also, observe that the error decreases like $1/N^2$ for fixed x in the interior of the interval $0 < x < \pi$.

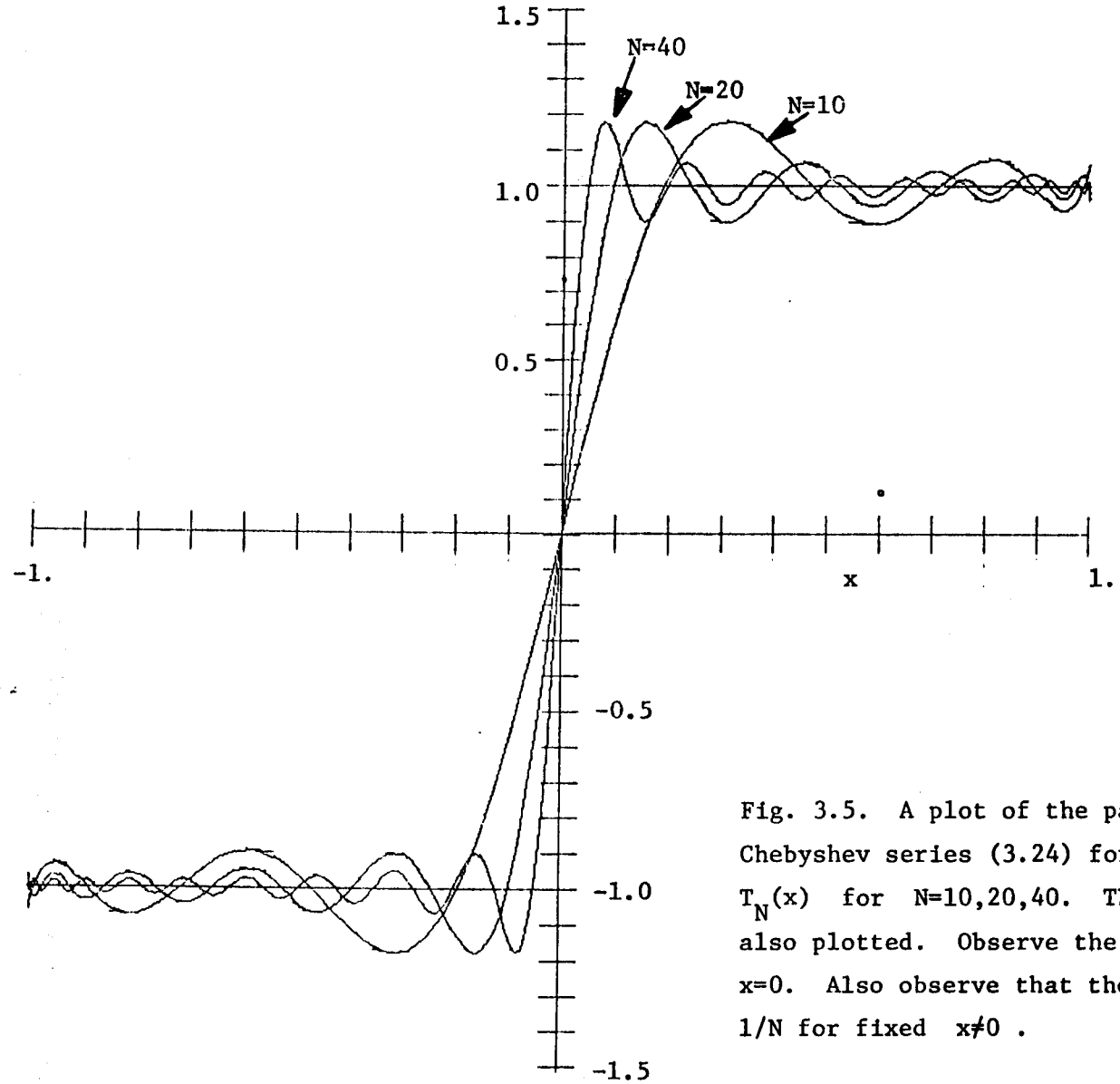

Fig. 3.5. A plot of the partial sums of the Chebyshev series (3.24) for sgn x truncated after $T_N(x)$ for N=10,20,40. The function sgn x is also plotted. Observe the Gibbs phenomenon near x=0. Also observe that the series converges like 1/N for fixed $x \neq 0$.

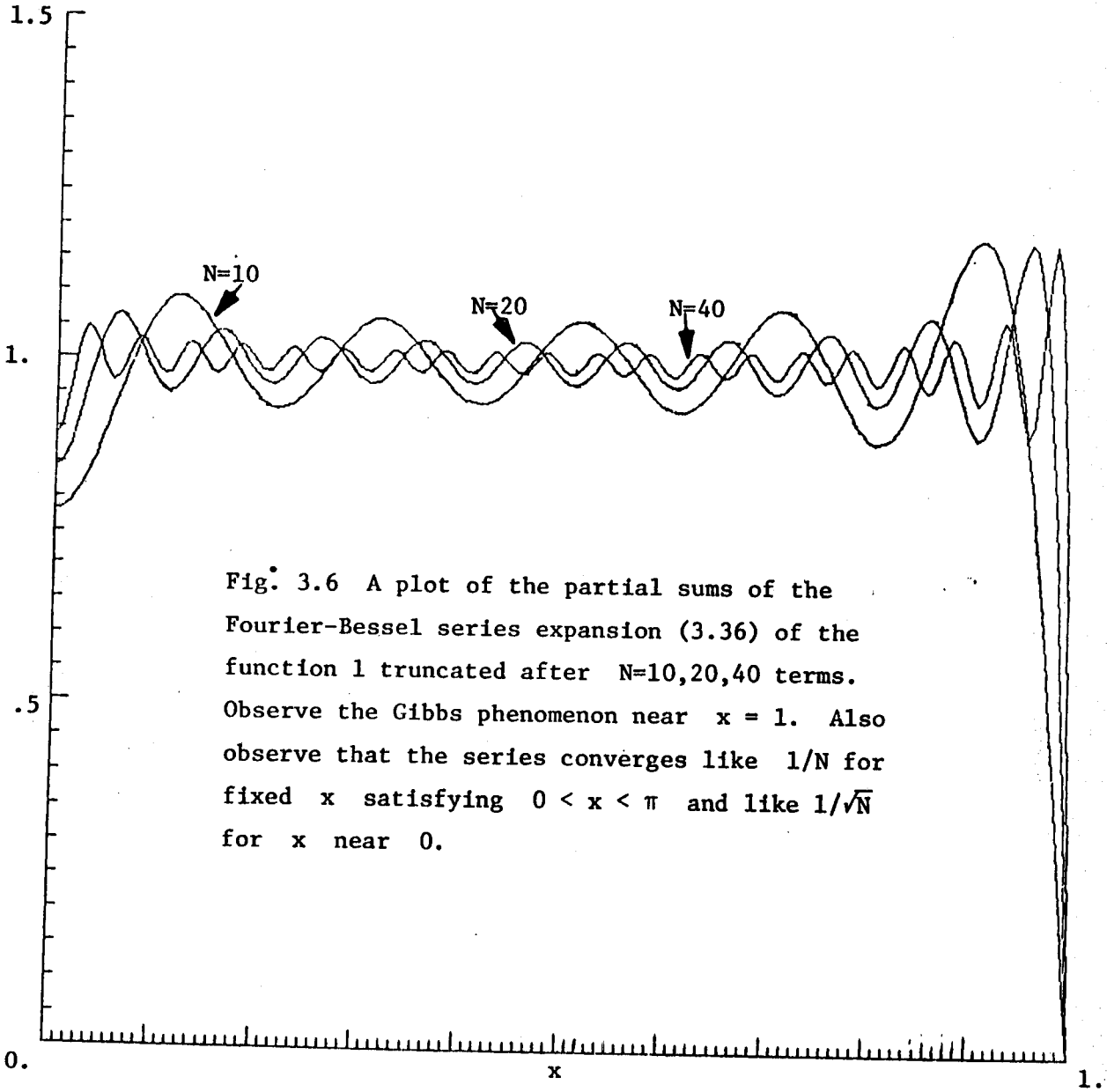

Fig. 3.6 A plot of the partial sums of the Fourier-Bessel series expansion (3.36) of the function 1 truncated after N=10,20,40 terms. Observe the Gibbs phenomenon near $x = 1$. Also observe that the series converges like $1/N$ for fixed x satisfying $0 < x < \pi$ and like $1/\sqrt{N}$ for x near 0.

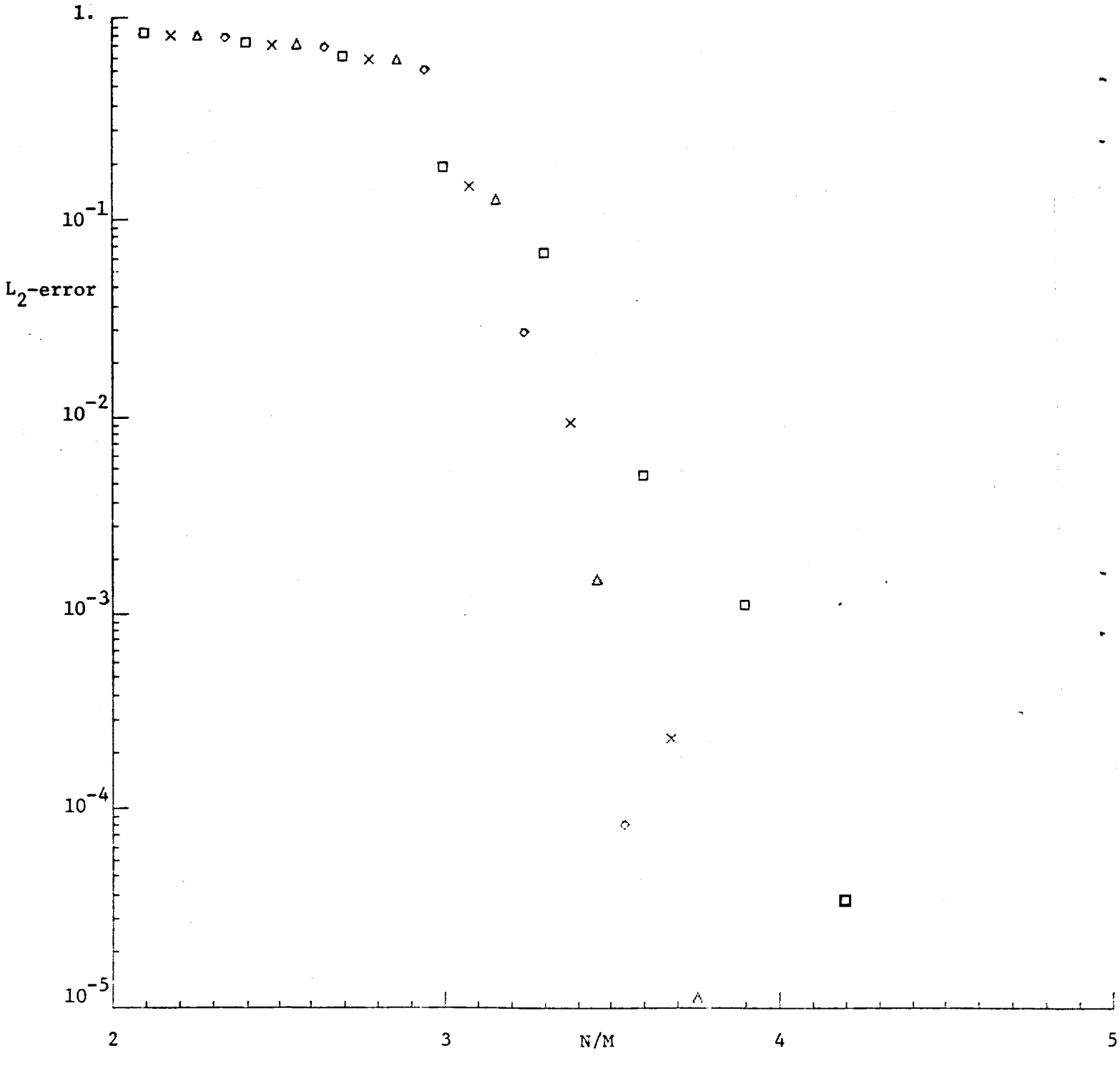

Fig. 3.7. A plot of the L_2-error in the Chebyshev series expansion (3.38) of $\sin(M\pi x)$ truncated after $T_N(x)$ versus N/M. The various symbols represent: □ N = 10; × N = 20; Δ N = 30; 0 N = 40. Observe that the L_2-error approaches zero rapidly when $N/M > \pi$.

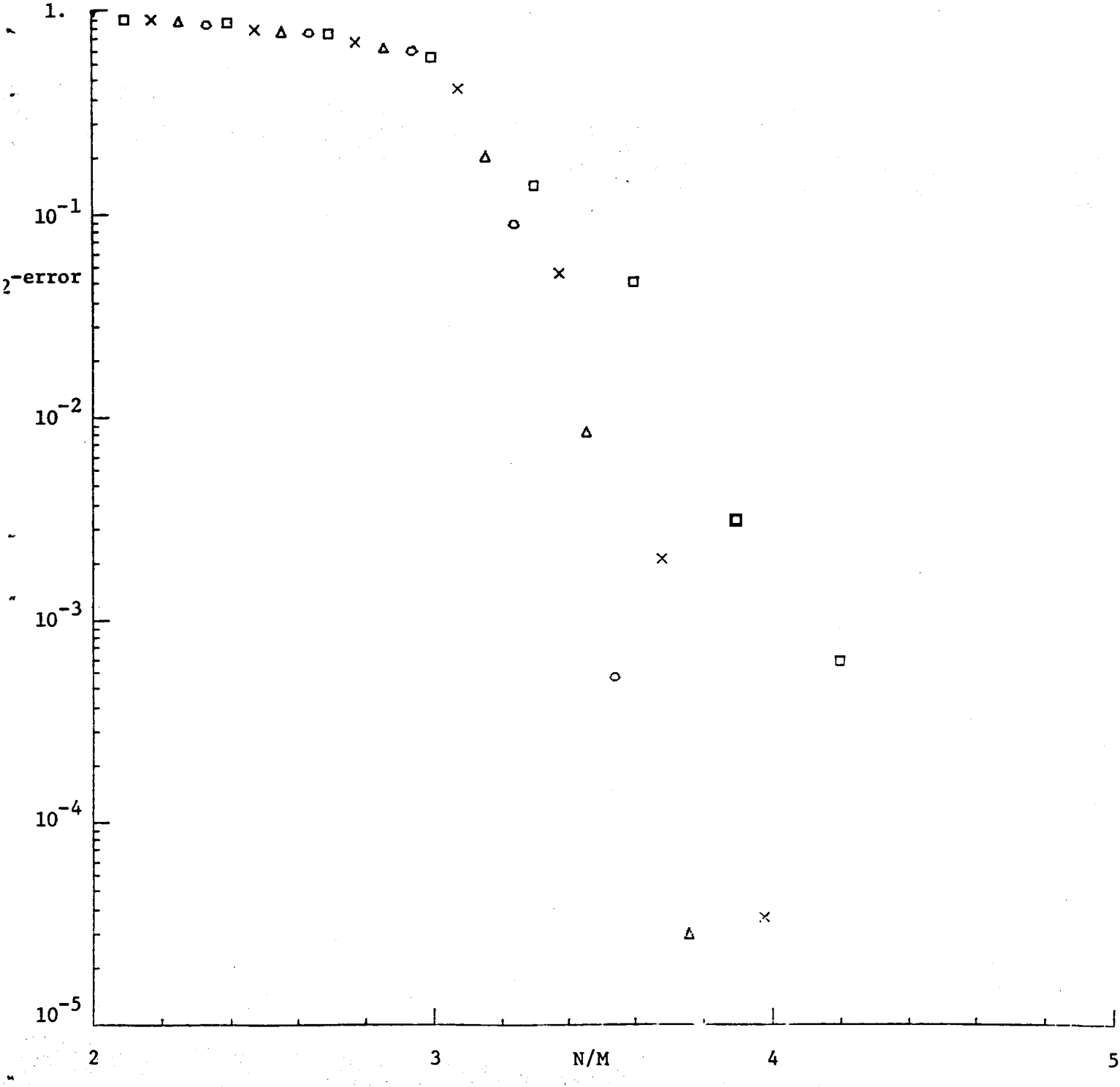

Fig. 3.8. A plot of the L_2-error in the Legendre series expansion (3.39) of $\sin(M\pi x)$ truncated after $P_N(x)$ versus N/M. The various symbols represent: □ N = 10; × N = 20; Δ N = 30; 0 N = 40. Observe that the L_2-error approaches zero rapidly when $N/M > \pi$.

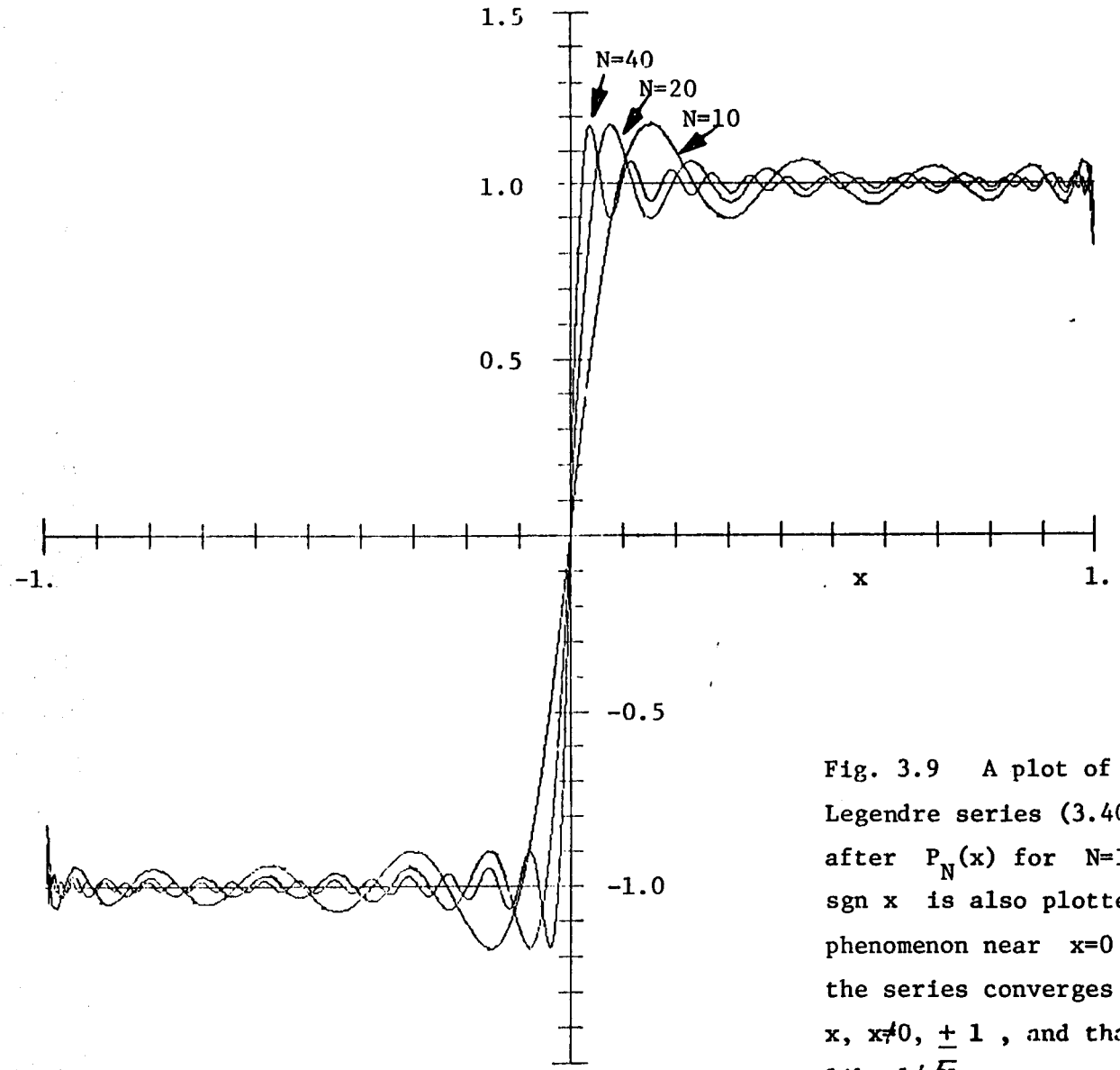

Fig. 3.9 A plot of the partial sums of the Legendre series (3.40) for sgn x truncated after $P_N(x)$ for N=10,20,40. The function sgn x is also plotted. Observe the Gibbs phenomenon near x=0 . Also observe that the series converges like 1/N for fixed x, $x \neq 0, \pm 1$, and that the series converges like $1/\sqrt{N}$ near $x = \pm 1$.

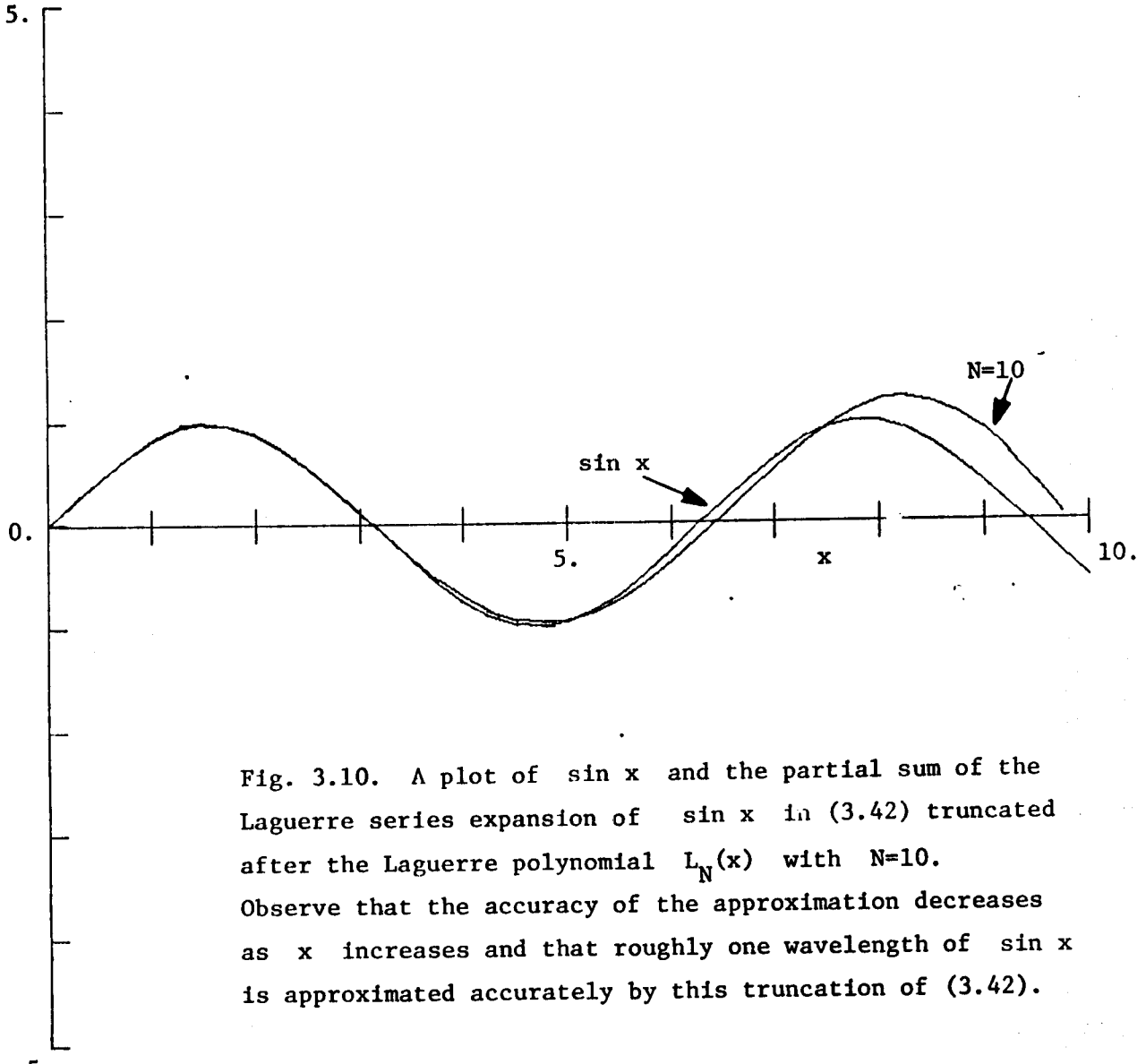

Fig. 3.10. A plot of $\sin x$ and the partial sum of the Laguerre series expansion of $\sin x$ in (3.42) truncated after the Laguerre polynomial $L_N(x)$ with $N=10$. Observe that the accuracy of the approximation decreases as x increases and that roughly one wavelength of $\sin x$ is approximated accurately by this truncation of (3.42).

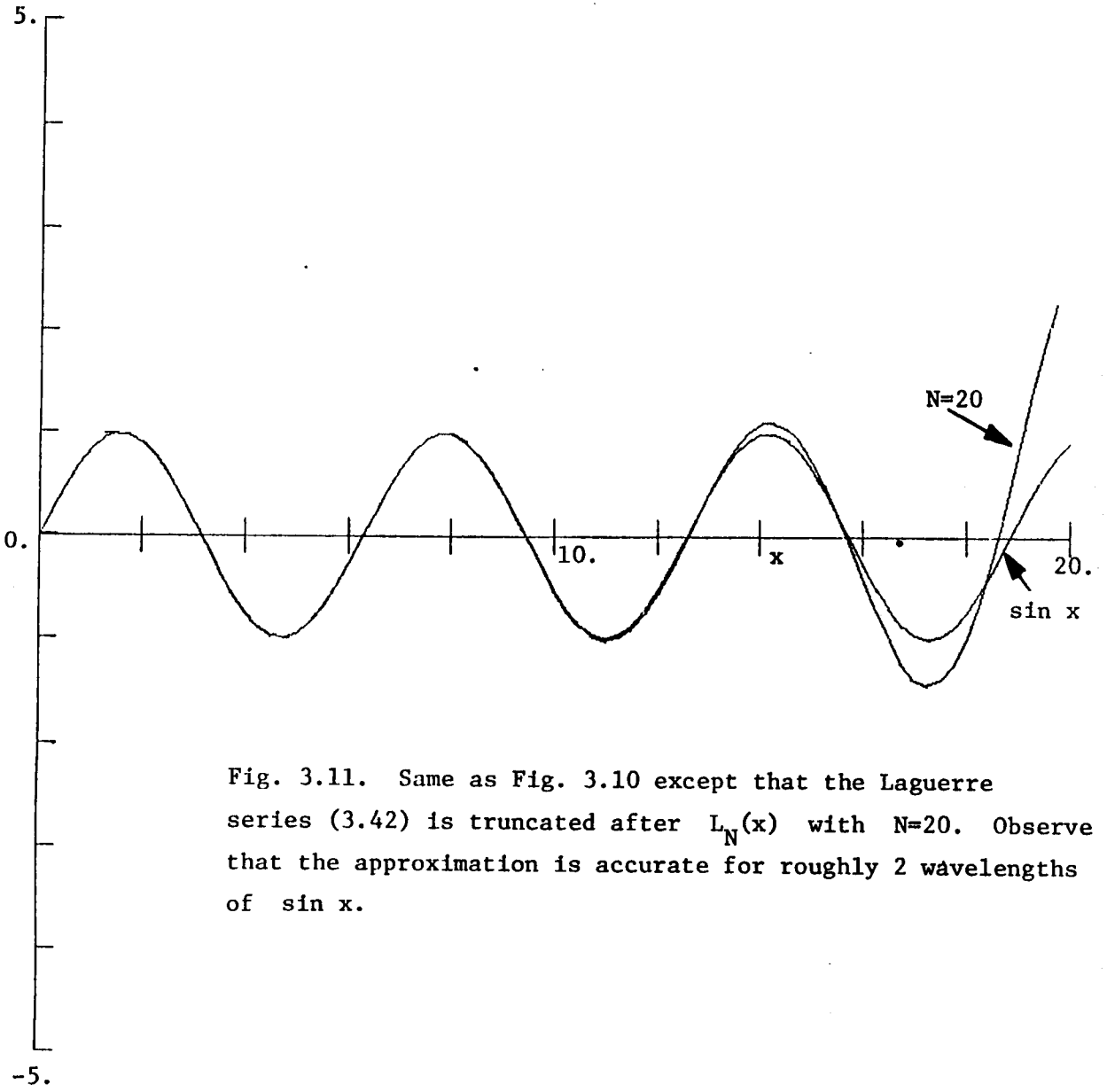

Fig. 3.11. Same as Fig. 3.10 except that the Laguerre series (3.42) is truncated after $L_N(x)$ with N=20. Observe that the approximation is accurate for roughly 2 wavelengths of sin x.

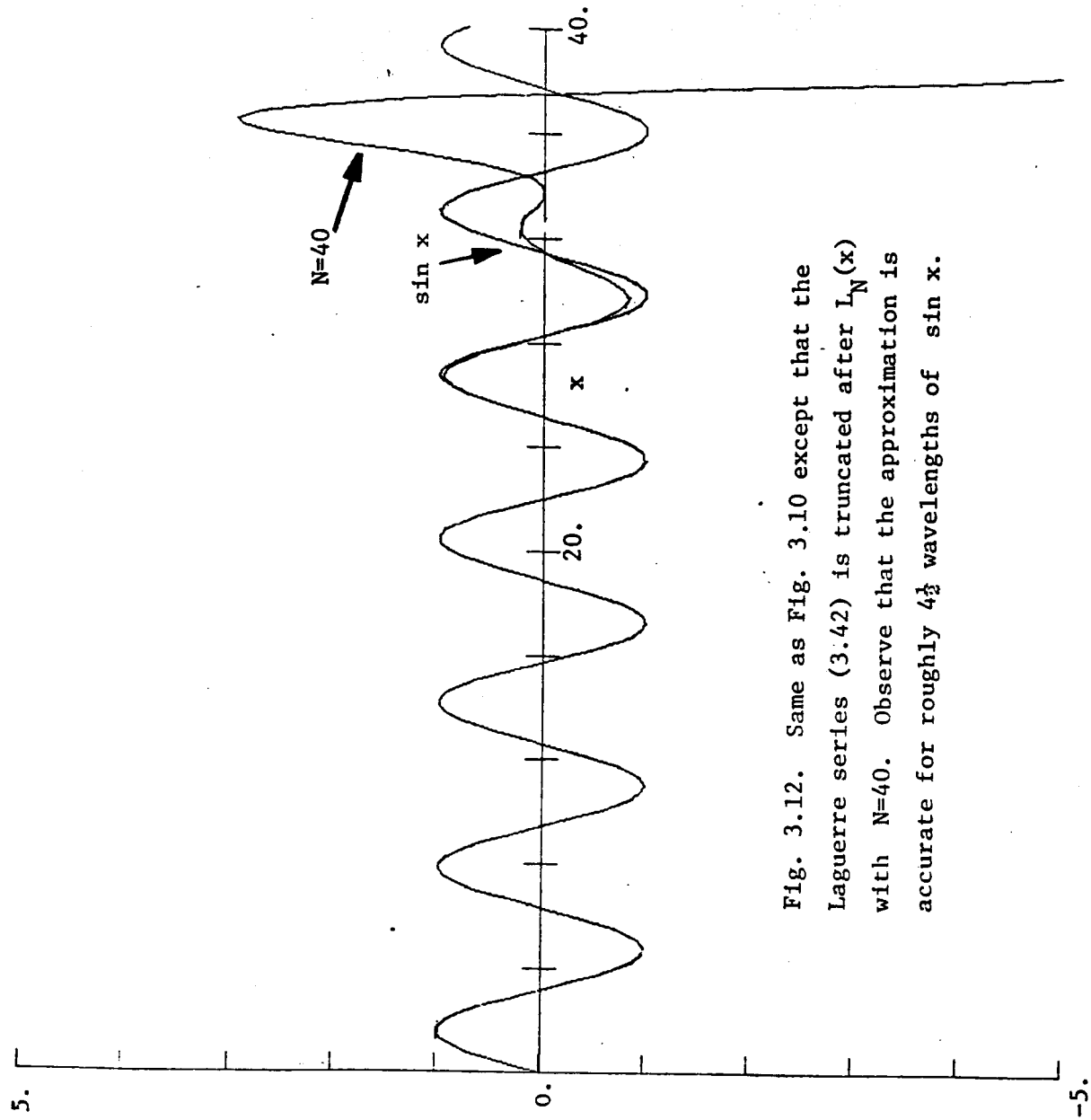

Fig. 3.12. Same as Fig. 3.10 except that the Laguerre series (3.42) is truncated after $L_N(x)$ with N=40. Observe that the approximation is accurate for roughly 4½ wavelengths of sin x.